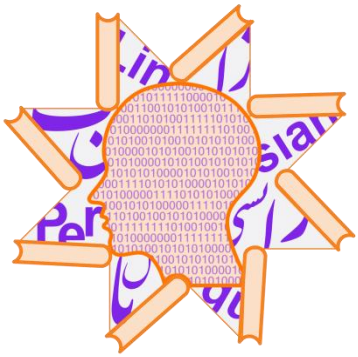
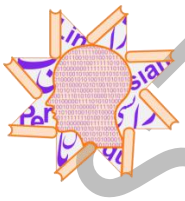


نخستین کنفرانس ملی

پژوهش‌های کاربردی در زبان‌شناسی رایانشی (با محوریت خط و زبان فارسی)

اسفند ۱۳۹۶



رئیس کنفرانس: دکتر محمدجواد دهقانی

قائم مقام رئیس: دکتر محمدرضا صالحی

دبیر علمی: دکتر محمدرضا فلاحتی قدیمی فومنی

دبیر اجرایی: دکتر محمدهادی فلاحی

طرح جلد: کریم فلاح

صفحه آرایی: کریم فلاح، اعظم دبستانی

شمارگان: ۳۰۰ نسخه

ناشر: اداره انتشارات مرکز منطقه‌ای اطلاع رسانی علوم و فناوری

تلفن: ۳۶۴۶۸۴۵۲-۰۷۱ نمابر: ۳۶۴۶۸۳۵۲-۰۷۱

مقالات این مجموعه در پایگاه استنادی علوم جهان اسلام (ISC) نمایه شده است.

پیش گفتار

زبان‌شناسی رایانشی حوزه‌ای میان رشته‌ای است که در سال‌های اخیر توجه بسیاری را به خود جلب کرده است. زبان‌شناسان، متخصصان علم اطلاعات و دانش شناسی، علوم رایانه، هوش مصنوعی، ریاضی، منطق، مردم شناسان و عصب شناسان از جمله متخصصانی هستند که به این رشته علاقه و توجه نشان داده‌اند. این رشته به بررسی و تولید الگوریتم‌ها و نرم‌افزارهای بررسی هوشمند داده‌های زبانی می‌پردازد. با توجه به گسترش روزافزون رایانه و علوم رایانه‌ای بررسی و پردازش و جایگاه بی‌نظیر آن در گسترش علوم، لازم است زبان فارسی نیز در این زمینه همگام با دیگر زبان های زنده‌ی دنیا نقش و جایگاه خود را حفظ و تقویت نماید.

با یاری خداوند متعال و به همت گروه زبان‌شناسی رایانشی مرکز منطقه‌ای اطلاع رسانی علوم وفن‌آوری نخستین کنفرانس ملی پژوهش‌های کاربردی در زبان‌شناسی رایانشی (با محوریت خط و زبان فارسی) در روزهای نهم و دهم اسفند ماه ۱۳۹۶برگزار گردید. در برگزاری این کنفرانس افراد و سازمان‌های مختلفی مشارکت و همکاری داشته‌اند که جا دارد سپاسگزار همه‌ی این عزیزان و حامیان باشیم. از آقای دکتر محمدجواد دهقانی، ریاست محترم مرکز منطقه‌ای اطلاع رسانی علوم وفن آوری و سرپرست پایگاه استنادی علوم جهان اسلام (ISC) و همچنین از جناب آقای دکتر محمد رضا صالحی، معاون محترم پژوهش و فناوری مرکز منطقه‌ای اطلاع رسانی علوم و فناوری به دلیل حمایت‌های بی‌دریغشان سپاس ویژه داریم. همچنین از اعضای محترم شورای علمی مرکز، اعضای محترم هیات علمی گروه‌های پژوهشی مرکز، روابط عمومی و همکاری‌های علمی بین المللی، اداره انتشارات، امور اداری و امورمالی و کلیه‌ی عزیزانی که ما را در برگزاری این کنفرانس یاری رساندند، تشکر می‌نماییم. از سازمان‌ها و دانشگاه‌های حامی کنفرانس از جمله یونسکو، پایگاه استنادی علوم جهان اسلام، دانشگاه اصفهان، دانشگاه صنعتی شریف، دانشگاه شیراز، دانشگاه علامه طباطبایی، انجمن زبان‌شناسی، انجمن ترویج زبان و ادب فارسی، پژوهشگاه علوم انسانی و بنیاد سعدی که به صورت علمی و معنوی حامی کنفرانس بودند، سپاسگزاریم.

اعضای هیات علمی کنفرانس

رئیس کنفرانس: دکتر محمد جواد دهقانی

قائم مقام رئیس: دکتر محمد رضا صالحی

دبیر علمی: دکتر محمدرضا فلاحتی قدیمی فومنی

دبیر اجرایی: دکتر محمدهادی فلاحی

اعضای هیات علمی کنفرانس:

دکتر مصطفی عاصی: استاد زبان‌شناسی پژوهشگاه علوم انسانی و مطالعات فرهنگی

دکتر جلال رحیمیان: استاد زبان‌شناسی دانشگاه شیراز

دکتر سیدعلی اصغر میرباقری فرد: استاد زبان و ادبیات فارسی دانشگاه اصفهان

دکتر سید آیت الله رزمجو: استاد زبان انگلیسی دانشگاه شیراز

دکتر ناصر رشیدی: استاد زبان انگلیسی دانشگاه شیراز

دکتر رحمان صحراگرد: استاد زبان‌شناسی کاربردی دانشگاه شیراز

دکتر رضا مراد صحرایی: دانشیار زبان‌شناسی دانشگاه علامه طباطبایی

دکتر فرخ حاجیانی: دانشیار فرهنگ و زبان‌های باستانی دانشگاه شیراز

دکتر علیرضا خرمایی: دانشیار زبان‌شناسی دانشگاه شیراز

دکتر سعید مهرپور: دانشیار زبان انگلیسی دانشگاه شیراز

دکتر مهرزاد منصوری: دانشیار زبان‌شناسی دانشگاه شیراز

دکتر محمدرضا قانع: دانشیار گروه علم اطلاعات و دانش‌شناسی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

دکتر محمد بحرانی: استادیار زبان‌شناسی رایانشی، دانشگاه صنعتی شریف

دکتر احسان چنگیزی: استادیار فرهنگ و زبان‌های باستانی دانشگاه علامه طباطبایی

دکتر امیر سعید مولودی: استادیار زبان‌شناسی دانشگاه شیراز

دکتر بهزاد مریدی: استادیار زبان‌شناسی دانشگاه پیام نور شیراز

دکتر هاجر صفاهیه: استادیار گروه علم اطلاعات و دانش شناسی مرکز منطقه‌ای اطلاع رسانی علوم و فناوری

دکتر بهاره پهلوان زاده: استادیار گروه طراحی و عملیات سیستم‌ها مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

دکتر علی گزنی: استادیار علم اطلاعات و دانش شناسی، مرکز منطقه‌ای اطلاع رسانی علوم و فناوری

دکتر حسن مقدس زاده: استادیار گروه علم اطلاعات و دانش شناسی مرکز منطقه‌ای اطلاع رسانی علوم و فناوری

دکتر حمید علیزاده: استادیار گروه زبان‌شناسی رایانه‌ای مرکز منطقه‌ای اطلاع رسانی علوم و فناوری

دکتر محمد باقر دستغیب: استادیار گروه طراحی و عملیات سیستم‌ها مرکز منطقه‌ای اطلاع رسانی علوم و فناوری

دکتر محمد رضا فلاحتی قدیمی فومنی: استادیار گروه زبان‌شناسی رایانه‌ای مرکز منطقه‌ای اطلاع رسانی علوم و فناوری

دکتر محمد هادی فلاحی: استادیار گروه زبان‌شناسی رایانه‌ای مرکز منطقه‌ای اطلاع رسانی علوم و فناوری

آقای شاپوررضا برنجیان: مربی گروه زبان‌شناسی رایانه‌ای مرکز منطقه‌ای اطلاع رسانی علوم و فناوری

حامیان علمی و معنوی کنفرانس

سازمان آموزشی، علمی و فرهنگی ملل متحد (یونسکو)

پایگاه استنادی علوم جهان اسلام (ISC)

دانشگاه شیراز

دانشگاه علامه طباطبائی

دانشگاه اصفهان

دانشگاه صنعتی شریف

پژوهشگاه علوم انسانی و مطالعات فرهنگی

انجمن ترویج زبان و ادب فارسی

انجمن زبان‌شناسی ایران

بنیاد سعدی

فهرست مقالات

مروری بر تحلیل احساسات با رویکرد لغتنامه، یادگیری ماشین و روش‌های ترکیبی

سمیرا السادات سجادی* و **حمید رستگاری****

چکیده

با تکامل سریع اینترنت و رشد رسانه‌های آنلاین مـا ننـد وب سـایت هـا، شـبکه‌هـای اجتماعی، وبلاگ ها، پورتال‌های آنلاین، افراد و سازمان‌هـا قـادر هسـتند تـا نظـرات و تجارب شخصی خود را بیان کنند. این نظرات برای افراد، سازندگان و فروشندگان کالا و خدمات بسیار مفید است و تحلیل احساسات مفید و ارزشمندی از جنبه ی مثبت و منفی یک موضوع مشخص را استخراج می‌کند. با توجه به حجم زیاد نظـرات و نقـدها در این رسانه‌ها سیستمی برای استخراج اطلاعات از وب احساس می‌شود. دو رویکـرد مبتنی برلغتنامه و یادگیری ماشین جهت استخراج خودکار معنایی و تحلیـل احسـاس استفاده می‌شود. رویکرد لغتنامـه جهـت طبقـه بنـدی نظـرات از دیکشـنری و پیکـره احساسی و رویکرد یادگیری ماشین از الگوریتم‌های طبقه بندی ماشین بهره می‌گیـرد. در این مقاله به تحلیل و ترکیـب دو رویکـرد فـوق پرداختـه شـده اسـت و روش‌هـای ترکیبی به دلیل پیوند چند طبقه بند و استفاده از روش‌هـای متاهیورسـتیکی دقـت و کارایی بیشتری نسبت به رویکردهای دیگر، ارائه کردند.

واژه‌های کلیدی: تحلیل احساس، روش مبتنی بر یادگیری ماشین، روش مبتنی بر لغتنامه، روش‌هـای ترکیبی

۱- مقدمه

زمینه‌های تحلیل احساسات در سال‌های اخیر توجه زیادی را به خود جلب کـرده اسـت. بـا رشد انفجاری متون تولید شده توسط کاربر در اینترنت، استخراج اطلاعات مفید به طور خودکار از اسناد فراوان، توسط متخصصان ضرورت یافته است. تعداد زیادی وبلاگ‌ها ، تالارهای گفتگـو، شبکه‌های اجتماعی وجود دارند که کاربران نظرات خود را در مورد کـالا و یـا خـدمات، افـراد و

* دانشکده مهندسی کامپیوتر، واحد نجف آباد، دانشگاه آزاد اسلامی، نجف آباد، ایران Sajadi2016s@gmail.com

** دانشکده مهندسی کامپیوتر، واحد نجف آباد، دانشگاه آزاد اسلامی، نجف آباد، ایران (نویسنده مسئول)
rastegari@iaun.ac.ir

موضوعات مختلف در آن درج می‌نمایند. به این ترتیب برای کسی که خواستار خرید کالا و آگاهی از یک موضوع خاص می‌باشد حجم زیادی از اطلاعات دروب وجود دارد. بسیاری از شرکت‌ها تمرکز کمپین‌های بازاریابی خود را بر تجزیه و تحلیل نظرات آنلاین قرار می‌دهند. با این حال، دستیابی به دانش مناسب از چنین حجمهای زیاد و تحلیل دستی این نظرها سخت و پرهزینه است. در اینجا بحث تحلیل احساسات ضرورت می‌یابد و با تحلیل خودکار نظرها، اطلاعات ارزشمندی که در آن نهفته است کشف می‌شود. تحلیل احساسات، علم بین رشته‌ای از داده کاوی، پردازش زبان طبیعی و متن کاوی است و می‌تواند احساسات بیان شده در متون را تعیین کند و مشخص نماید نظرات مثبت، منفی یا خنثی هستند. دو رویکرد اساسی برای تحلیل احساس شامل رویکردهای یادگیری ماشین و لغتنامه وجود دارد و تاکنون تحقیقات بسیاری در زمینه نظر کاوی و تحلیل احساسات با بهره از هر یک از رویکردها و پیوند آنها انجام گرفته است. در این مقاله ما به تحلیل و پژوهش رویکردهای موجود و ترکیب آنها خواهیم پرداخت. روش‌های ترکیبی به دلیل پیوند چند طبقه بند و استفاده از روش‌های متاهیورستیکی دقت و کارایی بیشتری نسبت به رویکردهای دیگر، ارائه کردند[۱۰, ۹].

۲- سطوح مختلف تجزیه و تحلیل احساسات

تجزیه و تحلیل احساسات می‌تواند به عنوان یک فرآیند طبقه بندی ۳ سطحی شامل: سطح سند، سطح جمله، سطح کلمه در نظر گرفته شود.

سطح سند: بیانگر احساس مثبت یا منفی در کل یک سند است. سطح جمله: مشخص کننده بیان مثبت، منفی یا خنثی است. منظور از خنثی یعنی جمله بیان کننده نظر نیست. این سطح به صورت نزدیکی با "طبقه بندی ذهنیت" وابستگی دارد. جملات ذهنی بیان کننده دید و نظر ذهنی هستند و ازجملات عینی که بیان کننده اطلاعات حقیقی هستند متمایز می‌شود [۳,۹].

سطح کلمه: دو سطح قبلی آنچه مردم دوست دارند و یا دوست ندارندرا کشف نمی نماید. و به جای نگرش به ساختار زبان (سند، پاراگراف، جمله) سطح کلمه مستقیما خود نظر را دنبال می‌کند. سطح کلمه به سطح ویژگی نیز شناخته شده است. مانند " اگر چه این رستوران سرویس چندان خوبی ارائه نمی کند ولی من هنوز به آن علاقه دارم" این نظر دید مثبت را

بیان می‌کند اما نمی توانیم بگوییم تماما مثبت است. درحقیقت این نظردر دید کلی مثبت است و در مورد ارائه سرویس دهی آن گرایش منفی است [۳,۵].

۳- مروری بر کارهای گذشته:

تحقیقات زیادی در زمینه تحلیل احساسات و تعین قطبیت اسناد صورت گرفته است. در این قسمت مرورکارهای پیشین بر اساس تکنیک‌های یادگیری ماشین و مبتنی برلغتنامه مورد بررسی قرار گرفته است. یادگیری ماشین به عنوان یکی ازشاخه‌های هوش مصنوعی، به تنظیم و اکتشاف شیوه‌ها و الگوریتم‌ها می‌پردازد که بر اساس آن رایانه‌ها توانایی تعلیم و یادگیری پیدا می‌کنند. در رویکرد مبتنی بر لغتنامه از لیستی از کلمات و عبارات احساسی استفاده می‌شود که لغتنامه احساسی نامیده می‌شود. از لغتنامه احساسی برای جهت گیری معنایی استفاده می‌شود[۵].

در شکل۱ نمودار کلی تکنیک‌های طبقه بندی احساسات نشان داده شده است.

۳-۱ یادگیری مبتنی بر ماشین

روش یادگیری ماشین به دو روش نظارت شده و نظارت نشده تقسیم می‌شود. موفقیت این دو روش به انتخاب ویژگی‌ها وابسته است. از معایب آن نیازمند بودن به کدهای انسانی جهت ایجاد مجموعه آموزش و صرف زمان و انجام محاسبات می‌توان اشاره کرد.

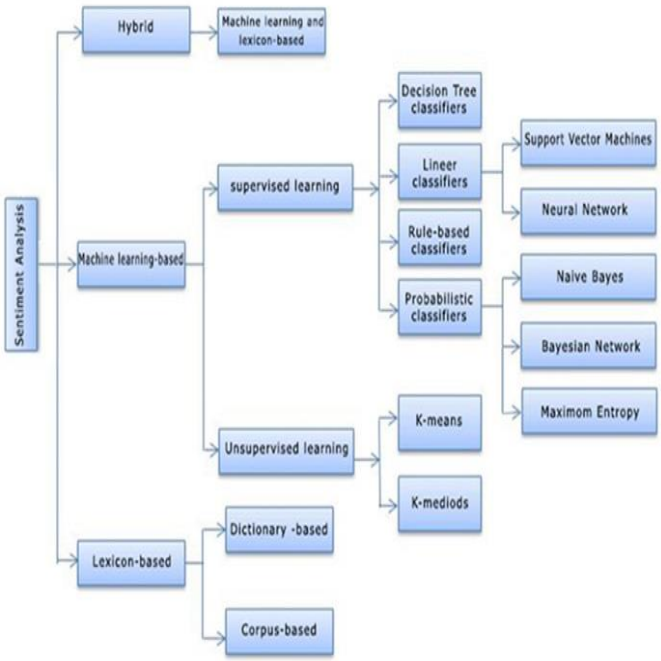
۳-۱-۱ یادگیری تحت نظارت

این الگوریتم‌ها با نظارت و برچسب عملیات دسته بندی را انجام می-دهند. در دسته بندی مبتنی بر یادگیری تحت نظارت، دو مجموعه اسناد مورد نیاز است: مجموعه آموزش و مجموعه تست. مجموعه آموزش جهت آموزش الگوریتم و مجموعه تست جهت ارزیابی هدف در نظر گرفته می‌شود. انواع روش‌های یادگیری تحت نظارت شامل:

۳-۱-۱-۱ طبقه بندی درخت تصمیم گیری

درخت تصمیم یکی از ابزارهای قوی و متداول برای دسته بندی و پیش بینی می‌باشد.

درخت تصمیم گیری یک ساختاردرختی است. در این ساختار هر گره داخلی آزمونی را بر روی یک ویژگی مشخص می‌کند. گره‌های برگ، کلاسها یا توزیع کلاسها را ارئه می‌نمایند. بالاترین گره در درخت، گره ریشه است. درتحقیق و پژوهش انجام شده در سال ۲۰۱۳ با استفاده از فرکانس سند معکوس و اهمیت کلمه یافت شده ویژگی‌های فیلم IMDb با استفاده از الگوریتم CART درکل سند استخراج شد و دقت طبقه بندی به دست آمده LVQ**۷۵**٪ می‌باشد[۱۷].



شکل ۱: تکنیک‌های طبقه بندی احساسات

۳-۱-۱-۲ طبقه بندی خطی

- **ماشین بردار پشتیبانی**[1]

مجموعه‌ای از نقاط در فضای n بعدی داده‌ها هستند که مرز دسته‌ها را مشخص می‌کنند و مرزبندی و دسته بندی داده‌ها بر اساس آنها انجام می‌شود. ماشین بردار پشتیبان، یک دسته بند یا مرزی است که با قرار دادن بردارهای پشتیبان، بهترین دسته بندی و تفکیک

1 Support Vector Machine

بین داده‌ها را برای ما مشخص می‌کند. نزدیکترین داده‌های آموزشی به آن صفحه‌های جداکننده، بردار پشتیبان نامیده می‌شود. پژوهش‌های انجام شده در زمینه ماشین بردار پشتیبان به صورت زیر می‌باشد:

پنگ و لی در سال ۲۰۰۲ به بررسی نظراف فیلم با استفاده از ویژگی‌های موقعیت (unigrams,)bigram , POS و الگوریتم‌های یادگیری ماشین(ماشین بردار پشتیبان، ماکزیمم آنتروپی ،نایو بیز) پرداختند. در این تحقیق کارایی بالای unigrams اثبات شد و الگوریتم‌های ماشین بردار پشتیبان و نایوبیز دقت بیشتری را نشان دادند. ٪۷۸,۷ =SVM ٪۸۲,۹ = Naive Bayes [۱۱].

پنگ ولی درسال ۲۰۰۴ نیز با بررسی نظراف فیلم با استفاده از الگوریتم SVM دقت ۸۶,۴۰ ٪ را کسب کردند [۱۲]. تحقیقات و پژوهش هایی نیز در سال ۲۰۱۱ با بررسی سایت آمازون و نظرات وبلاگ‌ها و مرور محصولات با استفاده الگوریتم ماشین بردار پشتیبان و ویژگی‌های مختلف n-grams انجام گرفت و دقت نتایج آن بدین شرح می‌باشد: سایت آمازون =۶۱ ٪ نظرات وبلاگ و مرور محصولات = ۹۱,۵۱٪ [۲,۱۴].

- **شبکه عصبی:**

شبکه عصبی شامل بسیاری از نورون‌ها است و نورون واحد اصلی آن است. ورودی‌های نورون توسط بردار خط X_i نشان داده شده است مجموعه‌ای از وزن‌های w وجود دارد که با هر نورون مورد استفاده قرار می‌گیرد تا محاسبات عملکرد ورودی‌ها را آن را انجام دهد و بر اساس ورودیها و وزنها خروجی تولید می‌شود.

۳-۱-۱-۳ طبقه بندی مبتنی بر قانون

در طبقه‌بندی‌های مبتنی بر قانون، فضای داده با مجموعه‌ای از قوانین مدل سازی می‌شود. در این نوع دسته بندی رکوردها توسط مجموعه‌ای از قوانین "IF …. Then" مشخص می‌شود. در سمت چپ مجموعه‌ای از قوانین و در سمت راست برچسب کلاس قرار می‌گیرد. تحقیق و پژوهش انجام شده در سال ۲۰۱۵ به تحلیل وبلاگ‌های چینی با استفاده از الگوریتم مبتنی بر قانون پرداخت و دقت ۷۵٪ را ارائه کرد[۹].

۳-۱-۱-۴ طبقه بندی احتمالی

- **نایوبیز**[1]

از قضیه بیز برای پیش بینی احتمال اینکه ویژگی مشخص شده متعلق بـه یـک برچسـب خاص باشد استفاده می‌شود. استدلال بیزی روشی بر پایه احتمالات برای استنتاج کردن است.

C : برچسب کلاس A : ویژگی ها

$$P(C|A) = \frac{P(A|C)*P(C)}{P(A)} \quad (۱)$$

$$P\ (C|\ A_1, A_2, A_3, \dots, A_n) \quad (۲)$$

هدف محاسبه یک دسته رکورد مفروض با مجموعه ویژگی‌هـای (A1,A2,A3,…,An) می‌باشد. درواقع از بین دسته‌های موجود به دنبال پیدا کردن دسته‌ای هستیم که مقـادیر A را بیشینه کند[۱۵] .

تحقیقاتی در سـال ۲۰۰۵ در ۴ حـوزه (کتـاب، فـیلم ، پایگـاه اطلاعـاتی وب و خـدمات پشتیبانی محصول) با استفاده از ویژگی‌های موقعیت

(trigrams, bigrams , unigrams) و الگوریتم نایوبیز انجـام گرفـت. ترکیـب تمـام ویژگی‌های n-grams به بهترین شکل عمل می‌کنند در حالی که در کل حوزه‌هـا unigrams، گاهیtrigrams و گاهی هم bigram بهتر عمل می‌کنند. و نتایج دقـت در ۴ حـوزه بـدین صورت مـی‌باشـد: فـیلم و کتـاب =۵۰٪ ، پایگـاه اطلاعـاتی وب = ۵۱٪ ، خـدمات پشـتیبانی محصول = ۵۲ ٪ [۱۳].

پژوهش هایی نیـز در سـال ۲۰۱۱ بـه بررسـی رسـتوران‌هـای آنلایـن بـا اسـتفاده از ویژگی‌های موقعیت (unigram، bigram، trigram) و

الگوریتم نایوبیز انجام شد و کارایی بهتر bigram به اثبـات رسـید و نتـایج دقـت ۹۳٪ کسب شد [۱۴].

در سال ۲۰۱۶ نیز پژوهشی در زمینه بررسی نظـرات فـیلم در تـوییتر بـا اسـتفاده از ویژگی unigram و الگوریتم‌های ماشین(ماشین بردار پشـتیبان و نـایو بیـز) انجـام گرفـت و

1 Naive Bayse

دقت ٪۷۵ : SVM و

٪۶۵ : Naive Bayes حاصل شد [۱].

- **شبکه‌های بیزین**

شبکه‌های بیزی، در واقع ترکیبی از دو شاخه نظریه گراف و نظریه احتمال هستند، این شبکه‌ها عمدتا نشان دهنده روابط علی و معلولی میان گراف معلوم باشد، مدلهای احتمالاتی میتوانند برای استدلال و پیش بینی در مورد متغیرها بکارروند و در صورت نامشخص بودن ساختار گراف، با استفاده از این مدلها میتوان به یادگیری ساختار مدل پرداخت و استدلال و پیشبینی در مورد متغیرها را انجام داد. پژوهش انجام شده در سال ۲۰۰۹ با استفاده از ویژگی موقعیت unigram و الگوریتم طبقه بندی بیزین به بررسی نظرات وبلاگ‌ها پرداخت و دقت ٪۹۱,۲۱ را ارائه نمود[۱۵].

- **ماکزیمم آنتروپی**[1]

طبقه بندی ماکزیمم آنتروپی یک طبقه بندی احتمالاتی است که متعلق به کلاس مدل‌های نمایشی است. این طبقه بندی می‌تواند مورد استفاده برای حل انواع مختلفی از مشکلات طبقه بندی متن مانند تشخیص زبان، طبقه بندی موضوع، تجزیه و تحلیل احساسات وغیره استفاده شود. علاوه بر این، حداکثر طبقه بندی آنتروپی زمانی استفاده می‌شود که ما نمی توانیم استقلال شرطی ویژگی‌ها را فرض کنیم. یکی از مشکلات طبقه بندی متن وابستگی بین کلمات متن می‌باشد که به وضوح مستقل از یکدیگرنیستند. و مطابق با فرمول زیر محاسبه می‌شود:

$$P_{ME}(c \mid d, \lambda) = \frac{\exp[\sum_i \lambda_i f_i(c,d)]}{\sum_c \exp[\sum_i \lambda_i f_i(c,d)]} \qquad (۳)$$

C کلاس، d کلمه و λ بردار وزنی است[۱۰] .

پژوهشی در سال ۲۰۱۰ با استفاده از ویژگی Dependency relationبه بررسی سایت آمازون پرداخت و نتایج ٪۷۲.۶ Precision: ,٪۷۸,۷ Recall: ٪۷۴.۵ =F حاصل

1 Maximum Entropy

گردید[۱۸]. در سال۲۰۱۴ نیز تحلیل داده‌های تویتر با استفاده از دو الگوریتم یادگیری ماشین (SVM وME) صورت گرفت و نتایجی با دقت ٪ ۹۵=ME و ٪۹۵=SVM ارائه شد[۹].

۳-۱-۲: یادگیری نظارت نشده

این الگوریتم‌ها بدون ناظر و برچسب عملیات دسته بندی را انجام می- دهند. انواع روش‌های یادگیری نظارت نشده شامل:

- الگوریتم K-means

روش (K-Means) یکی از روش‌های خوشه بندی بدون نظارت است. در این روش ابتدا نقاطی به صورت تصادفی انتخاب می‌شود. سپس داده‌ها با توجه با میزان نزدیکی (شباهت) به یکی از این خوشه‌ها نسبت داده می‌شوند و بدین ترتیب خوشه‌های جدیدی حاصل می‌شود. درهر تکرار با میانگین گیری از داده‌ها، مراکز جدیدی برای آنها محاسبه می‌شود و مجددا داده‌ها رابه خوشه‌های جدید نسبت می- دهیم. این روند تا زمانی ادامه پیدا می‌کند که دیگر تغییری در داده‌ها حاصل نشود.

پژوهشی در سال ۲۰۱۰ با استفاده از انتخاب ویژگی TF-IDF [1] و خوشه بندی با استفاده از الگوریتم k-means برای تحلیل نظرات فیلم انجام گرفت و دقت ٪۷۸ حاصل شد[۲].

- **الگوریتم K-mediods**

این الگوریتم مشابه الگوریتم k-means است. در الگوریتم k-means مراکز خوشه ها براساس میانگین عناصر داخل خوشه محاسبه می‌شود اما در الگوریتم K-mediods براساس محاسبه میانه عناصر خوشه می- باشد. تحقیقاتی در سال ۲۰۱۷ با استفاده ازالگوریتم k-mediods و تکنیک‌های خلاصه سازی متن، به تحلیل نظرات آنلاین هتل‌ها پرداخت و اطلاعات جامعی درمورد جنبه‌های مثبت و منفی نظرات ارائه کرد[۸].

1 Term Frequency-Inverse Document Frequency

۳-۲ رویکرد مبتنی بر لغتنامه

مجموعه‌ای از کلمات و اصطلاحات شناخته شده که برای ارتباط سنتی طراحی شده اند و حاوی بار احساسی می‌باشند. مانند Opinion Finder lexicon . رویکرد مبتنی بر لغتنامه شامل روش مبتنی بر لغتنامه و روش مبتنی بر پیکره می‌باشد.

۳-۲-۱ روش مبتنی برلغتنامه

مجموعه‌ای از عبارات معروف که به صورت دستی با جهت‌های شناخته شده جمع آوری می‌شود. این مجموعه با جستجوی مترادف -ها و متضادها قابل ا فزایش هستند. برای نمونه لغتنامه WordNet با مترادف‌ها و متضادها گسترش یافت و SentiWordNet به وجود آمده است. قطبیت کلمات داخل متن با استفاده از دیکشنری‌های احساسی مشخص می‌شود. و از معایب لغتنامه ناتوانی در پیدا کردن کلمات احساسی با دامنه و جهت خاص می‌باشد. پژوهشی در سال ۲۰۱۴ از روش یادگیری بدون نظارت مبتنی بردیکشنری با استفاده از سطح ویژگی و لغت نامه WordNet به تعیین جهت معنایی جملات پرداخت و دقت ۷۴٪ را ارائه نمود[۱۶].

۳-۲-۲ روش مبتنی بر پیکره

رویکرد مبتنی بر منبع، هدف ارائه واژه نامه هایی است که مربوط به یک دامنه خاص می‌باشد. این واژه نامه‌ها از مجموعه‌ای از اصطلاحات عینی ساخته شده است که از طریق جستجوی کلمات مرتبط با استفاده از تکنیک‌های آماری و معنایی گسترش میابند. پژوهشی در سال ۲۰۰۷ با انتخاب ویژگی Graph distance measurement و روش مبتنی بر پیکره به تحلیل نظرات داخل وبلاگ‌ها پرداخت و دقت ٪۸۲,۷ تا ٪۹۵,۷ حاصل گردید [۲].

۳-۳ ترکیب الگوریتم‌های یادگیری ماشین و لغتنامه

مطالعات زیادی در استفاده از روش‌های یادگیری ماشین بر روی طبقه بندی معنایی وجود دارد. در سال‌های اخیراستفاده از روش‌های ترکیب که نتایج چندین طبقه بند را با هم ترکیب می‌کند رو به افزایش است. نتایج به دست آمده از پیوند چندطبقه بند دقت بیشتری

نسبت به نتایج یک طبقه بند دارد. پژوهشی در سال ۲۰۱۳ با استفاده از ویژگی‌های موقعیت (trigram، bigram،unigram) و روش ترکیب واژگانی و الگوریتم ماشین(SVM-PSO) که یک روش ترکیب متاهیوریستیکی می‌باشد انجام گرفت و دقت نتایج آن به این صورت می‌باشد.

جدول ۱: نتایج الگوریتم (SVM-PSO ,SVM)

SVM	SVM-PSO
Accuracy = ٪۷۱,۸۷	Accuracy=٪۷۷,۰۰
Precision=٪۶۸,۸۱	Precision=٪۷۷,۵۶
Recall=٪۸۱,۸۷	Recall=٪۷۶,۱۳

از نتایج به دست آمده روش ترکیبی SVM-PSO دارای دقت محاسباتی بیشتری نسبت به SVM **می**باشد [۴]. در سال ۲۰۱۵ نیز پژوهشی در سطح ویژگی و ترکیب روش واژگانی و الگوریتم ماشین SVM انجام گرفت ودقت٪۷۸ حاصل شد[۶]. تحقیقاتی در سال ۲۰۱۷با استفاده از روش جستجوی ترکیبی فاخته و الگوریتم k-means که یک روش متاهیوریستیکی می‌باشد و با استفاده از بردار ویژگی، در ۴ مجموعه داده مختلف از پیام‌های توییتر انجام گرفت و نتایجی با دقت بیش از ٪۶۷,۴۵ حاصل شد[۷].

۴. نتیجه گیری:

تجزیه و تحلیل احساسات به زمینه تحقیقاتی بسیار محبوب تبدیل شده است و در این زمینه تحقیقات فراوانی صورت گرفته است اگر چه تکنیک ها و الگوریتم‌های مورد استفاده برای تجزیه و تحلیل احساسات به سرعت پیشرفت می‌کنند، با این حال، بسیاری از مشکلات در این زمینه مانند تشخیص(کنایه‌ها ،ضرب المثل‌ها و نظرات مخرب) هنوز حل نشده است. در این مقاله ما به تحلیل و پژوهش احساسات با رویکر مبتنی بر دیکشنری و یادگیری ماشین و ترکیب آنها پرداختیم. از میان روش‌های موجود، روش‌های ترکیبی به دلیل پیوند چند طبقه بند و استفاده از روش‌های متاهیورستیکی دقت و کارایی بیشتری نسبت به تک طبقه بندها،

ارائه کردند.

منابع

[1] Ravi, K., & Vadlamani, R. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89, 14- 46.

[2] Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment nalysis and opinion mining: A survey. *International Journal*, *2*(6), 282-292.

[3] Alhojely, S. (2016). Sentiment analysis and opinion mining: A survey. *International Journal of Computer Applications*, 150, 1-4.

[4] Basari, A.S.H., et al. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, 53, 453-462.

[5] Liu, B. (2012). Sentiment analysis and opinionmining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

[6] Bhadane, C., Dalal, H., & Doshi (2015). Sentiment analysis: Measuring opinions. *Procedia Computer Science*, *45*, 808-814.

[7] Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, *53*(4), 64-779.

[8] Hu, Y-H., Chen, Y-L., & Chou, H. L. (2017). Opinion mining from online hotel reviews – A text summarization approach. *Information Processing & Management*, *53*(2), 436-449.

[9] Pradhan, V. M., Vala, J., & Balani, P. (2016). A survey on sentiment analysis algorithms for opinion mining. *International Journal of Computer Applications*, *133*(9), 1-5

[10] Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: A survey of techniques, *arXiv preprint arXiv:1601.06971*,139, 1-11.

[11] Pang, B. (2002). Thumbs up?: Sentiment lassification using machine learning techniques. In Proceedings of *the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Pang, B., Lee, L., & Vaithyanathan, S. *Association for Computational Linguistics*, 10, 79-86.

[12] Pang, B. (2004). A sentimental education: Sentimen analysis using subjectivity summarization based on minimum cuts. In Proceedings of

The 42nd Annual Meeting on Association for Computational Linguistics, Pang, B., Lee, L., *Association fo Computational Linguistics*, 1-8.

[13] Aue, A., & Gamon, M. (2005). Sentiment classifiers to new domains: A case study. InProceedings of *Recent Advances in Natural Language Processing (RANLP)*, 1-7.

[14] Govindarajan, M., & Romina, M. (2013). A survey of classification methods and applications for sentiment analysis. *The International Journal of Engineering And Science (IJES)*, *2*(12), 11-15.

[15] Errano-Guerrero, J. et al. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, *3*(11), 18-38.

[16] Sharma, R., Nigam, S., & Jain, R. (2014). Polarity detection at sentence level. *International Journal of Computer Applications*, *86*(11), 1-5.

[17] Jotheeswaran, J., & Kumaraswamy, Y. S. (2013). Opinion mining using decision tree based feature selection through Manhattan hierarchical cluster measure. *Journal of Applied Information Technology*, *58*(1), 1-9.

[18] Safdar, M., & Khan, M. J. I. (2015). Opinion mining for customer feedback: A survey. *International Journal of Scientific Research and Engineering Studies*, *2*(2), 1-4.

تحلیل متن کاوی گرایش‌های پژوهشی حوزه ترجمه ماشینی

حمید علیزاده زوج*

چکیده

این پژوهش با هدف شناسایی گرایش‌های پژوهشی اصلی حوزه ترجمه ماشینی انجام شد. برای شناسایی این گرایشها از روش متن کاوی استفاده شد. ترجمه ماشینی حوزه‌ای میان رشته است که هدف آن ترجمه متون علمی در یک زبان با متن مشابه در زبانی دیگر است. در این پژوهش انتشارات پژوهشی این حوزه در سالهای ۲۰۰۶ تا ۲۰۱۵ به روش متن کاوی مورد تحلیل قرار گرفت. پیکره پژوهش از انتشارات علمی این حوزه در پایگاه اسکوپوس استخراج گردید. متن کاوی با استفاده از روش هم رخدادی واژه‌ها انجام شد. این روش بینشی کامل نسبت به ساختار علمی و گرایش‌های پژوهشی این حوزه ارایه نمود. تحلیل هم رخدادی واژه‌ها یکی از روش‌های تحلیل محتوا است که در این پژوهش روی پیکره انتشارات حوزه ترجمه ماشینی اجرا شد. از دیداری سازی حوزه‌های دانش،برای نمایش شبکه واژگان عناوین و چکیده‌های مقالات ترجمه ماشینی استفاده شد.در نهایت نتایج تحلیل متن کاوی، ساختار علمی این حوزه را بر اساس خوشه‌های موضوعی آن شناسایی کرد. یافته‌های پژوهش شامل شناسایی چهار خوشه موضوعی است که گرایش‌های اصلی تحقیقاتی در این حوزه بشمار می‌آیند. در این میان دو خوشه برتر تحت عنوان ترجمه ماشینی آماری و ترجمه خودکار و ارتباطات کاربران به عنوان گرایش‌های اصلی پژوهشی حوزه ترجمه ماشینی معرفی شد.

واژه‌های کلیدی: ترجمه ماشینی، متن کاوی، ساختار علم، تحلیل هم رخدادی واژگان

۱- مقدمه

با افزایش بهره گیری از امکاناتی که فناوری اطلاعات و ارتباطات در اختیار محققین قرار داده است، فرایندهای تولید، ذخیره سازی و اشاعه اطلاعات دچار تغییرات شگرفی شده است. امروزه پایگاه‌های اطلاعاتی عظیم میلیونها رکورد اطلاعاتی را در اختیار دارند. علاوه بر این شبکه جهانی وب نیز سرشار از منابع متنی است. به جرات می‌توان گفت کمتر موضوعی را

* استادیار گروه زبانشناسی رایانه‌ای، مرکز منطقه‌ای اطلاع رسانی علوم و فناوری، alizade1377@gmail.com

می‌توان یافت که جستجوی اطلاعات در وب، حجم عظیمی از منابع را در آن خصوص بازیابی ننماید.

این حجم عظیم منابع متنی باعث شده است که تحلیل و پردازش متون به صورت دستی در محیط اطلاعاتی امروز کاری صعب و طاقت فرسا تلقی شود. نتایج بازیابی شده نیز دیگر چندان قابل اعتماد و تعمیم نیست. استفاده از سیستم‌های خودکار تحلیل متن، می‌تواند جوابگوی این نیاز باشد. این دستاوردها به نحو شگفت انگیزی در وقت و انرژی کاربر صرفه جویی می‌نماید.

از جمله نظام‌های خودکاری که در این زمینه به یاری کاربر شتافته تا بهره مناسبی از منابع سایر زبان‌ها ببرد، نظام ترجمه ماشینی است. ترجمه ماشینی زیر شاخه‌ای از حوزه پردازش زبان طبیعی است. علوم دیگر از جمله فناوری اطلاعات و زبانشناسی نیز در تحقیقات آن مشارکت دارند. هدف ترجمه ماشینی، ترجمه زبان‌های انسانی به شکلی خودکار است که کاربر را قادر سازد از این زبان‌ها و منابع نوشتاری آنها بهره ببرد.

حوزه‌های درگیر در توسعه دانش ترجمه ماشینی دارای تنوع بسیار می‌باشد. در عین حال جامعه پژوهشی بزرگی نیز در سراسر دنیا به پژوهش در این حوزه و انتشار یافته‌های آن از طریق مجلات علمی و یا ارائه در کنفرانس‌های بین المللی مشغولند.

حجم زیاد پژوهش‌ها و ابزار ترجمه ماشینی، نیاز به معیارهای ارزیابی در این حوزه را ضروری ساخته است[۱]. از جمله حوزه‌های پژوهشی مهم ترجمه ماشینی می‌توان به بررسی کارامدی این نظام‌ها در اجرای دقیق ترجمه متون اشاره نمود. بدیهی است که علاوه بر سرعت، کیفیت ترجمه ماشینی از نکات حائز اهمیت بشمار می‌رود. این افزایش کیفیت در طول زمان به مدد ارزیابی نظام‌های موجود و مشخص شدن نقاط ضعف آنها حاصل شده است. پژوهش در این حوزه با شتاب و وسعت مثال زدنی به پیش می‌رود. جنبه‌های اقتصادی ترجمه ماشینی و نیاز به معیارهای جدید بررسی کارامدی، از حوزه‌های پژوهشی دیگری است که به موارد پیشین اضافه شده است. بنابراین نیازی ضروری برای بررسی وضعیت تولید علم و پژوهش در این حوزه و پیش بینی سمت و سوی تحقیقات آینده در آن احساس می‌شود.

در جهت آشنایی با وضعیت کنونی تحقیق در این حوزه باید به شناسایی نقاط اصلی پژوهشی این حوزه پرداخت. با شناسایی آنها می‌توان به تحقیقات جدید دست زد. باید دید

جامعی نسبت به گذشته، حال و پیش بینی آینده این حوزه داشت. چنین چشم اندازی با تحلیل متن کاوی انتشارات حوزه ترجمه ماشینی حاصل می‌شود. بنابراین مساله اصلی این پژوهش شناسایی حوزه‌های موضوعی اصلی و چالش‌های تحقیقاتی این حوزه است.

۲- مروری بر پژوهش‌های مرتبط

ترجمه ماشینی استفاده از سیستم‌های رایانه‌ای برای ترجمه متن است. دو زبان منبع و هدف در این فرایند دخیل می‌باشد[۲]. هدف این فرایند آن است که مشارکت انسانی در آن به حداقل برسد. رویکردهای مختلفی در این حوزه وجود دارد که جدیدترین و موفق ترین آنها شیوه ترجمه ماشینی آماری است. دراین رویکرد برای ترجمه از پیکره‌های متنی موازی استفاده می‌شود. این شیوه، برخلاف دیگر رویکردها چندان وابستگی به دانش انسانی ندارد.

امروزه در بسیاری از رشته‌های علمی از تحقیقات علم سنجی برای شناسایی موثرترین مولفین، مقالات و نشریات یک حوزه استفاده شده است. در رشته زبانشناسی رایانه‌ای و حوزه ترجمه ماشینی اما چنین پژوهش هایی کمتر انجام شده است. از معدود پژوهش‌های انجام شده در این خصوص می‌توان به پژوهشی تحت عنوان از زبانشناسی رایانه‌ای تا تاریخ نگاری الگوریتمیک اشاره کرد[۳]. در این پژوهش برخی مشکلات ترجمه ماشینی ذکر شده است. ضمنا بر اهمیت نگاشت معنایی واژه‌ها در پژوهش‌های علم سنجی تاکید ویژه شده است.

در پژوهشی دیگر[۴] به بررسی علم سنجی حوزه ترجمه ماشینی پرداخته شده است. روش تحلیل کمی و تحلیل استنادی مقالات این حوزه نشان داد که ترجمه ماشینی به شاخص هایی علاوه بر کارامدی احتیاج دارد. این پژوهش از جمله اولین گام‌ها در تحلیل علم سنجی ترجمه ماشینی است. در آن تحلیل مجموعه مقالات یک کنفرانس آمریکایی در حوزه زبانشناسی با شاخص‌های علم سنجی تحلیل شد. همانگونه که مشخص است این تحلیل به دلیل محدودیت پیکره مورد بررسی قابل اعتماد و تعمیم نمی باشد.

با بررسی پژوهش‌های داخل کشور، پژوهشهایی که در آن از روش علم سنجی یا متن کاوی برای بررسی تحقیقات ترجمه ماشینی استفاده شده باشد کمترمشاهده می‌شود. از موارد محدود مرتبط با این زمینه می‌توان به پژوهش‌های زیر اشاره کرد:

فلاحتی، احمدی نسب و خانی(در دست چاپ) به بررسی روند مطالعات ترجمه در ایران

پرداختند. آنها با بررسی نشریات رتبه دار وزارت علوم تلاش نمودند تا حوزه‌های پژوهشی برتر و کمبودهای تحقیقاتی در این زمینه را مشخص سازند. در نهایت موضوعاتی چون اصول و روش ترجمه، آموزش ترجمه و ترجمه متون مذهبی از جمله مهم ترین حوزه‌های پژوهشی در این مجموعه شناخته شد. تفاوت پژوهش حاضر با این پژوهش در گستره بین المللی آن و استفاده از روش تحلیل علم سنجی و متن کاوی حوزه‌های دانش است.

در پژوهشی دیگر مجموعه مقالات دو دهه اخیر کنفرانس‌های زبانشناسی با روش علم سنجی تحلیل شد[۵]. مقالات هفت سمینار زبان‌شناسی داخلی با معیارها و شاخص‌های علم سنجی تحلیل شد. یافته‌های این پژوهش نشان داد که در طول زمان چه حوزه هایی بیشتر مورد استقبال قرار گرفته است. موثر ترین مقالات و مولفین و دانشگاهها مشخص شد و افزایش سهم زنان مولف نشان داده شد.

بهره گیری از روش متن کاوی برای تحلیل ساختارعلمی و گرایش‌های پژوهشی در حوزه‌های مختلف به کرات انجام شده است. متن کاوی به روش‌های مختلف انجام می‌شود. یکی از شیوه‌های مرسوم، استفاده از تحلیل هم رخدادی واژه‌ها است. این شیوه اولین بار توسط محققان فرانسوی بکار گرفته شد[۶]. از آن زمان پتانسیل‌ها و امکانات این روش با بهره گیری از تکنیک‌ها و ابزار جدید توسعه بیشتری یافته است. برخی پژوهشگران تحلیل هم واژه را نوعی تحلیل کتابسنجی دانسته اند، که در آن هم رخدادی واژه‌ها مورد بررسی قرار می‌گیرد[۷]. در این روش، پویایی رشته‌های علمی با استفاده از ماتریس هم رخدادی واژه‌ها که از پیکره‌های انتشارات علمی استخراج شده است، متن کاوی و بازنمون می‌گردد.

از این روش در تحلیل موضوعی بسیاری از رشته‌های علمی استفاده شده است. کارهایی در حوزه نانوتکنولوژی[۸]، مدیریت دانش[۹]،فنوم انسانی[۱۰]، ژنتیک[۱۱] و اطلاعات پزشکی[۱۲] از این جمله است. این روش تاکنون در حوزه ترجمه ماشینی استفاده نشده است و این پژوهش در نوع خود اولین بار صورت می‌پذیرد.

۳- روش شناسی

این پژوهش به روش متن کاوی انجام شده است. در روش متن کاوی واژه‌های کلیدی انتشارات حوزه ترجمه ماشینی از عناوین و چکیده مقالات پیکره پژوهش استخراج گردید.

جامعه پژوهش، کلیه مقالات منتشره این حوزه در سطح بین المللی در سالهای ۲۰۰۶ تا ۲۰۱۵ است. پیکره پژوهش با جستجوی عبارت ترجمه ماشینی و ترجمه خودکار در فیلد عنوان و چکیده مقالات پایگاه اسکوپوس استخراج شد. آنگاه با استفاده از فیلتر مقاله، تنها مقالات نشریات و کنفرانس‌های علمی مورد بررسی قرار گرفت. اطلاعات مورد نیاز برای تحلیل متن کاوی از این پیکره استخراج گردید. برای تحلیل متن کاوی، کلید واژه‌های عناوین و چکیده مقالات موجود در پیکره استخراج شد. سپس کلیدواژه‌های استخراج شده با نرم افزار ووسویور نرمال سازی شد. جهت تولید ماتریس هم رخدادی واژه‌ها که نشان دهنده گرایشات موضوعی حوزه ترجمه ماشینی است از الگوریتم دیداری سازی مشابهت‌ها استفاده شد. پس از تشکیل ماتریس هم رخدادی از نرم افزار ووسویور جهت دیداری سازی شبکه واژه‌ها استفاده شد. با تشکیل شبکه واژه‌ها از روش‌های تحلیل شبکه، نگاره‌های تولید شده تفسیر شد. بر این اساس خوشه‌های موضوعی تشکیل شده و روابط بین آنها بعنوان بازنمون متن کاوی پیکره پژوهش معرفی شد. این خوشه‌ها نشان دهنده گرایشات پژوهشی برتر حوزه ترجمه ماشینی است که در بخش بعد به تفصیل شرح داده می‌شود.

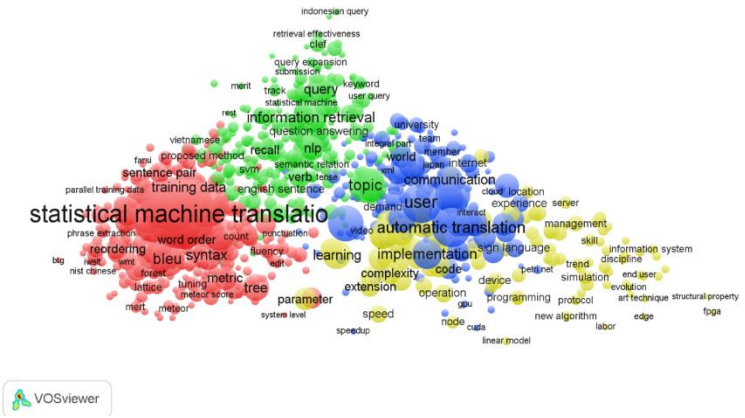
۴- یافته‌ها و بحث

تعداد انتشارات حوزه ترجمه ماشینی در سالهای ۲۰۰۶ تا ۲۰۱۵ که در پایگاه استنادی اسکوپوس نمایه شده است بالغ بر ۵۰۰۱ پیشینه است. از این تعداد ۲۹۸۹ مورد مقاله کنفرانس و ۱۰۸۵ پیشینه از مقالات چاپ شده در نشریات علمی و مابقی از سایر انواع مدارک است. از سوی دیگر، بررسی سال‌های مختلف دوره ده ساله این پژوهش نشان داد که این حوزه در سالهای مختلف از استقبال نسبی محققین بهره برده است.

در این سالها مولفین برترحوزه ترجمه ماشینی به این شرح است. وی با ۷۵ اثر، سومیتا با ۷۲ و نی با ۵۸ اثر رتبه‌های اول تا سوم را به خود اختصاص داده اند. نکته جالب در این میان وجود تعداد زیادی محقق با خاستگاه جنوب شرق آسیاست که نهادینه شدن پژوهش در این حوزه را در جغرافیای آن منطقه نشان می‌دهد. اگرچه بسیاری از این محققین هم اکنون جذب سازمان‌های علمی اروپا و آمریکا شده اند و از آن وابستگی سازمانی استفاده می‌کنند ولی این نشان دهنده ماهیت سیال جریان علم است.

بررسی عناوین نشریات و کنفرانس هایی که انتشارات حوزه ترجمه ماشینی را منتشر می‌کنند، نشانگر آن است که این حوزه ماهیتی بین رشته‌ای دارد و محققین با وابستگی‌های موضوعی متنوعی در این حوزه نقش ایفا می‌کنند. چنانکه بیشترین انتشارات ترجمه ماشینی خاستگاه علوم رایانه‌ای داشته و ۳۹۷۱ اثر در این حوزه قرار می‌گیرد. هنر و علوم انسانی با ۱۴۴۹ ، علوم اجتماعی با ۱۱۴۰، ریاضیات با ۸۷۱ ومهندسی با ۶۹۳ مقاله در رتبه‌های بعدی قرار دارد.

در جهت تحلیل متن کاوی، واژه‌های موجود در عنوان و چکیده مقالات ترجمه ماشینی استخراج و پس از انجام پردازش زبان طبیعی و حذف واژه‌های تهی، هم رخدادی واژه‌های موضوعی این رشته و شدت آن محاسبه گردید. شبکه حاصل از محاسبه هم رخدادی واژه‌ها برای ایجاد بینش در خصوص ساختارعلمی و نحوه تکامل این حوزه دیداری سازی شد که شکل مربوط به آن در ادامه ارایه شده است.



شکل۱. نمای شبکه‌ای از ساختار علمی حوزه ترجمه ماشینی

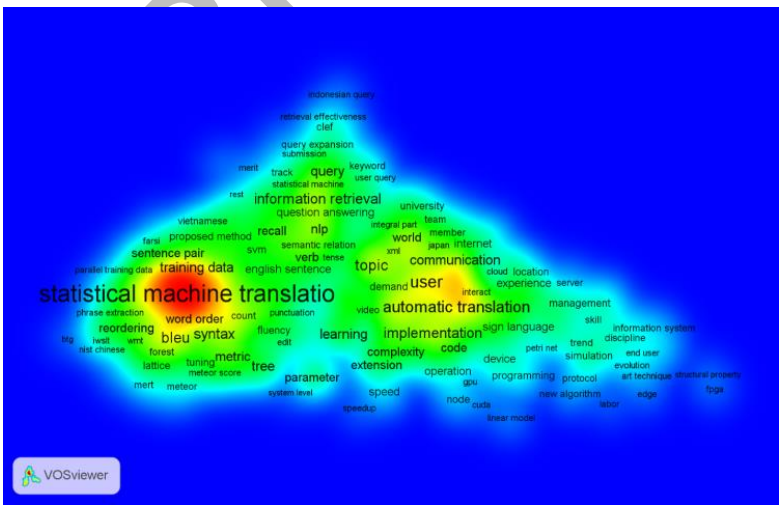
در این شکل ۴ خوشه موضوعی که ساختار کلی تحقیقات در حوزه ترجمه ماشینی را نشان می‌دهد به تصویر کشیده شده است. خوشه موضوعی با نام ترجمه ماشینی آماری با بیشترین تناوب و تراکم(سمت چپ شکل) نشان می‌دهد که رویکرد آماری در طراحی ماشین‌های ترجمه در حال حاظر رویکرد اصلی در تحقیقات این حوزه است. اگر چه وجود واژه

هایی چون نحو و نظم واژگان و درخت واژگان در این خوشه نشان دهنده یکپارچگی این تحقیقات با اصول زبانشناسی رایانشی است.

خوشه موضوعی که در وسط و رو به بالای شکل قرار دارد خوشه بازیابی اطلاعات است. وجود این خوشه در این شکل بیهوده نیست زیرا بازیابی اطلاعات بویژه در مبحث بازیابی بین زبانی وابستگی مستقیم با ترجمه ماشینی دارد. وجود اصطلاحاتی نظیر پرس و جو، تاپیک و سیستم پرسش و پاسخ موید شکل گیری هوشمندانه این خوشه و موفقیت رویکرد استفاده شده در این پژوهش است.

خوشه راست بالای این شکل با نام ترجمه خودکار و دارا بودن اصطلاحاتی نظیر کاربر و ارتباطات، نشان دهنده اهمیت ترجمه ماشینی در تعاملات انسانی و ارتباطات بین المللی است. در واقع این خوشه وجه علوم انسانی و جامعه شناسی را در تحقیقات این حوزه نمایندگی می‌کند.

خوشه پایین راست در این شبکه برخلاف خوشه‌های دیگر از تراکم درونی خاصی برخوردار نیست. بر اساس واژه‌های موضوعی موجود در این خوشه آن را خوشه طراحی و ارزیابی ترجمه ماشینی نامگذاری می‌کنیم. بطور کلی شکل حاضر نشان دهنده نحوه تکامل تحقیقاتی رشته بوده و نقشه راه تحقیقات آینده در این حوزه را نیز مشخص می‌کند.



شکل ۲. نمای تراکمی حوزه‌های داغ پژوهشی ترجمه ماشینی

در شکل ۲ نمای تراکمی حوزه‌های پژوهشی ترجمه ماشینی تشان داده شده است. در این شکل حوزه تحقیقات ترجمه ماشینی آماری با رنگ قرمز، داغ ترین حوزه پژوهشی این حوزه است. تعاملات کاربر و ترجمه ماشینی و نقش آن در ارتباطات با رنگ زرد دومین حوزه مورد علاقه برای پژوهش بشمار می‌رود. پس از آن سایر حوزه هایی که با رنگ سبز مشخص شده اند از اهمیت تقریبا یکسانی برخوردارند. موضوعاتی نیز در حاشیه آبی قرار گرفته اند که کمتر از سایر حوزه‌های اشاره شده مورد توجه قرار گرفته است.

۵- نتیجه گیری

در این پژوهش تلاش شد تا با تحلیل متن کاوی انتشارات ترجمه ماشینی، ساختار علمی تحقیقات این حوزه شناسایی گردد. برای تحلیل ساختار علمی و گرایش‌های موضوعی یک حوزه از روش‌های مختلفی استفاده می‌شود. روش هم استنادی و هم رخدادی واژه‌ها از روش‌های مرسوم این تحلیل بشمار می‌رود. در این پژوهش برای متن کاوی انتشارات ترجمه ماشینی از تحلیل هم رخدادی واژه‌ها استفاده شد. میزان بروندادهای علمی این حوزه نشان داد که در سالهای مختلف دوره مورد بررسی اقبال نسبتا مناسبی از تحقیقات ترجمه ماشینی وجود داشته است. در خصوص مشارکت رشته‌های مختلف در تحقیقات ترجمه ماشینی، این نکته قابل توجه است که جامعه تحقیقاتی متشکل از حوزه‌های مختلف در این تحقیقات مشارکت دارند. در واقع نیاز به ارتباطات انسانی با سایر اقوام و ملل موجب تقویت ترجمه شد. این امر نهادینه شدن تحقیقات آن در جامعه شناسی و علوم انسانی را به دنبال داشت. با ظهور فناوری اطلاعات حوزه‌های فنی نیز به این جرگه وارد شدند و به کمک زبانشناسی رایانه‌ای، ترجمه ماشینی شکل گرفت. این نتایج ماهیت بین رشته‌ای تحقیقات ترجمه ماشینی را نشان می‌دهد.

تحلیل متن کاوی ساختار و تکامل علمی حوزه ترجمه ماشینی، با نمایش دو بازنمون شبکه‌ای و تراکمی حوزه‌های تحقیقاتی ترجمه ماشینی به شناسایی گرایش‌های پژوهشی عمده این حوزه منجر شد. ساختار موجود با نمایش خوشه‌های موضوعی و موضوعات داغ پژوهشی می‌تواند پژوهشگران را در انتخاب مسایل پژوهشی یاری نماید. در هر رشته علمی گرایش‌های پژوهشی مختلفی وجود دارد. برخی از آنها نسبت به سایر گرایشها نقش موتور محرکه پژوهش در آن رشته را ایفا می‌کنند. خوشه‌های موضوعی که در شبکه نشان داده شد این نقش را به

عهده دارند. شبکه‌های علمی می‌توانند با خلاصه سازی انبوه داده‌ها بینشی عمیق تولید نمایند. محققان نیز می‌توانند با مشاهده گرایش‌های برتر تحقیقاتی در این شبکه، چالش‌های اصلی پژوهش در رشته را شناسایی کنند. در نهایت انجام پژوهش هایی از این دست می‌تواند با شناسایی نقاط مهم پژوهشی، به توسعه نظام‌های کارآمدتر ترجمه ماشینی منجر گردد.

منابع

[1] Voß, S. and Zhao, X., 2005. "Some Steps Towards a Scientometric Analysis of Publications in Machine Translation". In *Artificial Intelligence and Applications* (pp. 651-655).

[2] Hutchins, J., 2005. "The history of machine translation in a nutshell". Retrieved December, 20:2009, 2005.

[3] Garfield, E. 2001. "From computational linguistics to algorithmic historiography '. In *Symposium in Honor of Casimir Borkowski at the University of Pittsburgh School of Information Sciences*.

[4] Voß, S., & Zhao, X. 2005. "Some Steps Towards a Scientometric Analysis of Publications in Machine Translation". In *Artificial Intelligence and Applications* (pp. 651-655).

[5] gholi, f., 2014. "A survey of two decades of Iranian linguistics conference: a scientometric analysis". Zabzn shenakht.

[6] Rip, a., & Courtial, J.-P., 1984." Co-word maps of biotechnology: An example of cognitive scientometrics". *Scientometrics*, *6*(6), 381–400.

[7] Moed, H.F., Glänzel, W. and Schmoch, U., 2004. "Editors' introduction". In *Handbook of quantitative science and technology research* (pp. 1-15). Springer, Dordrecht.

[8] de Miranda Santo, M., Coelho, G.M., dos Santos, D.M. and Fellows Filho, L., 2006. "Text mining as a valuable tool in foresight exercises: A study on nanotechnology". *Technological Forecasting and Social Change*, *73*(8), pp.1013-1027.

[9] Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, Y. and Rajman, M., 1998. "Knowledge Management: A Text Mining Approach". In *PAKM* (Vol. 98, p. 9).

[10] Van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G. and Leunissen, J.A., 2006. "A text-mining analysis of the human phenome". *European journal of human genetics*, *14*(5), p.535.

[11] Krallinger, M., Valencia, A. and Hirschman, L., 2008." Linking genes to literature: text mining, information extraction, and retrieval applications for biology". *Genome biology*, *9*(2), p.S8.

[12] Holzinger, A., Geierhofer, R., Mödritscher, F. and Tatzl, R., 2008." Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses". *J. UCS*, *14*(22), pp.3781-3795.

طراحی و پیاده سازی سامانه ژورنال یاب با بهره گیری از سامانه‌های توصیه گر ترکیبی

محمدباقر دستغیب[*]، سارا کلینی[**]، بهاره پهلوان‌زاده[***] و امین زارع[****]

چکیده

یافتن نشریه مناسب برای ارسال یک مقاله یکی از مهمترین مراحل در جریان انتشار مقاله است. برای اکثر نویسندگان این کار به سادگی امکان پذیر نمی باشد، زیرا بسیاری از نشریات دارای موضوعات با تنوع بسیار گسترده‌ای هستند و بسیاری از مقالات را می‌توان مقالات بین رشته‌ای دانست که شامل چندین رشته تخصصی بوده و می‌توان در نشریات مختلفی در حوزه‌های مربوطه به چاپ رساند. لذا یافتن نشریات مطابق با موضوع مقاله به عنوان یکی از چالش‌ها در امر چاپ مقاله مطرح می‌باشد.در این مقاله، ویژگی‌های سامانه ژورنال یاب مرکز منطقه‌ای اطلاع رسانی علوم و فناوری[1]، به عنوان یک سامانه جامع جهت یافتن نشریه مورد نظر کاربرشرح داده می‌شود. در سامانه ژورنال یاب همه‌ی حوزه‌های مهم علمی و بیش از ۱۵۰۰ نشریه بررسی شده تا به نویسندگان کمک نماید تا نشریات مربوط به حوزه تخصصی مقاله را جهت ارسال مقاله خود پیدا کنند. در این تحقیق از الگوریتم ترکیبی سامانه‌های توصیه‌گر که یکی از شاخه‌های یادگیری توسط ماشین است استفاده شده است. ضریب همبستگی پیرسون را برای داده‌های آزمون و پاسخ‌هایی که سامانه ژورنال یاب برای مجموعه‌ی آزمون بدست آورده نشان می‌دهد که سامانه‌ی توصیه‌گر ژورنال یاب با همبستگی بالا (۰,۸۱) با جواب‌های واقعی توانسته است پاسخ صحیح را توصیه کند. همچنین با استفاده از معیار میانگین امتیاز متقابل نیز آزمون انجام شده و نهایتا میانگین امتیاز متقابل برای سامانه‌ی توصیه‌گر ژورنال یاب برابر ۰,۵۳ بدست آمد. بنابراین، نتایج مکتسب نشان می‌دهد که سامانه ژورنال یاب می‌تواند از نظر دقت پاسخ‌های ارائه شده و همچنین رتبه‌ی پاسخ‌های مرتبط امتیاز قابل قبولی را کسب نماید.

---

* مرکز منطقه ای اطلاع رسانی علوم و فناوری، dastgheib@ricest.ac.ir

** مرکز منطقه ای اطلاع رسانی علوم و فناوری، koleini@ricest.ac.ir

*** مرکز منطقه ای اطلاع رسانی علوم و فناوری، pahlevanzadeh@ricest.ac.ir

**** مرکز منطقه ای اطلاع رسانی علوم و فناوری، azare@ricest.ac.ir

1 http://journalfinder.ricest.ac.ir

واژه‌های کلیدی: سامانه‌های توصیه گر، زبان فارسی، ژورنال یاب، بازیابی اطلاعات، یادگیری ماشین.

۱. مقدمه

در عصری که با انفجار بیش از حد اطلاعات مواجه هستیم، از استراتژی‌های مختلفی برای تصمیم گیری‌های گوناگون از جمله انتخاب‌های انجام شده در حوزه علمی مانند انتخاب نشریه مورد مطالعه استفاده می‌شود. سامانه‌های توصیه گر برخی از این استراتژی‌ها را با هدف ارائه توصیه‌های خودکار که مقرون به صرفه و با کیفیت بالا باشد را ارائه می‌دهند. سامانه‌های توصیه گر برنامه‌های کامپیوتری هستند که "بهترین انتخاب" را به کاربران در زمینه‌های مختلف توصیه می‌کنند. رویکردهای متفاوتی جهت توسعه سامانه توصیه گر پیشرفته وجود دارد. از جمله می‌توان به رویکردهای الگوریتمی موجود جهت ارائه پیشنهادات خرید شخصی مانند فیلترینگ مبتنی بر محتوا، و همچنین روشهای همکاری و دانش محور اشاره کرد. الگوریتم‌های طراحی شده برای توصیف و ارائه پیشنهادات به کاربران به یک امر چالش برانگیز در برنامه‌های کاربردی وب تبدیل شده است. مسئله اصلی، رتبه بندی موارد بازیابی شده بر اساس پاسخ کاربران به منظور بهینه سازی سامانه است.

به طور کلی، در الگوریتم یک سامانه توصیه گر موارد زیر در نظر گرفته می‌شود[1]:

درک محتوا و فیلتر کردن آن: باید از تکنیک هایی برای فیلتر کردن محتوای کم کیفیت از استخر داده های موجود استفاده کرد. توصیه به محتوای کم کیفیت باعث آسیب به سامانه توصیه گر می‌شود. در هر سامانه توصیه گر، تعریف کم کیفیت، بستگی به ویژگی آن سامانه دارد.

مدل سازی پروفایل کاربر: باید پروفایل‌های کاربری ایجاد کرد که منعکس کننده مواردی باشد که کاربران احتمالا مورد استفاده قرار می‌دهند. این پروفایل‌ها می‌توانند براساس جمعیت شناسی، اطلاعات ارائه شده هویت کاربر در زمان ثبت، اطلاعات شبکه های اجتماعی یا اطلاعات رفتاری در مورد کاربران باشد.

نمره دهی: بر اساس پروفایل کاربری و پروفایل اقلام، باید تابع امتیازدهی برای تخمین احتمال مقادیر و ارزشهای آینده برای نمایش اقلام به کاربر با توجه به زمینه مورد علاقه وی بوجود آورد.

رتبه بندی: در نهایت، نیاز به مکانیسمی برای انتخاب یک لیست رتبه بندی شده از اقلام در جهت توصیه نمودن آنها به کاربر است. در ساده ترین سناریو، رتبه بندی ممکن است شامل موارد مرتب سازی بر اساس یک امتیاز واحد برای هر مورد باشد.

به طور خلاصه الگوریتم فوق را می‌توان به شرح زیر بیان نمود. سیگنال‌های ورودی بر اساس اطلاعات کاربر، اطلاعات آیتم مورد نظر و تاریخچه داده‌های تعاملی کاربر- آیتم، توسط مدل‌های آماری یادگیری ماشین برای تولید نمرات استفاده شده که میزان وابستگی کاربران به اقلام را تعیین می‌کند. نمره‌ها توسط ماژول رتبه بندی ترکیب شده تا یک لیست مرتب شده‌ای از اقلام را بر اساس ترتیب نزولی از اولویت به دست آمده تولید نماید.

در این قسمت به دسته بندی کلاسیک سامانه‌های توصیه گر پرداخته می‌شود:

سامانه‌های توصیه گر مبتنی بر محتوا[1]: در این سامانه شباهت اقلام بر اساس ویژگی‌های مرتبط با موارد مقایسه شده محاسبه می‌شود .در واقع، فرآیند اصلی انجام شده توسط یک توصیه گر مبتنی بر محتوا، شامل تطبیق ویژگی‌های پروفایل کاربر است که در آن موارد مورد علاقه کاربر با ویژگی‌های محتوای یک شیء (آیتم)، به منظور توصیه موارد جدید به کاربر ذخیره می‌شود[2].

فیلترینگ همکاری[2] : این روش به عنوان ساده ترین و اصلی ترین پیاده سازی‌های سامانه‌های توصیه گر محسوب شده که در آن به کاربر جاری و فعال مواردی را که سایر کاربران با سلیقه‌های مشابه در گذشته علاقه داشته‌اند را پیشنهاد می‌دهد[3]. بر خلاف سامانه-های توصیه گر مبتنی بر محتوا، فیلترینگ‌های همکاری می‌توانند مواردی با محتوای بسیار متفاوت را در صورتیکه پیش از این سایر کاربران علاقه‌ای به این آیتم‌های مختلف نشان داده باشند، توصیه کنند[4].

سامانه‌های محدودیت گرا[3] : این سامانه ها، به عنوان نوع دیگری از سامانه‌های توصیه گر دانش محور محسوب می‌شوند. از نظر دانش مورد استفاده، هر دو سامانه مشابه می‌باشند، تفاوت عمده این دو سامانه در راه حل محاسبه است[5].

---

1 Content-based
2 Collaborative filtering
3 Constraint-based systems

سامانه‌های توصیه گر مبتنی بر مورد[1] : در این سامانه‌ها که نوع دیگری از سامانه‌های دانش محور شناخته می‌شود، توصیه‌های مبتنی بر معیارهای تشابه مشخص می‌گردد. در حالی که در سامانه توصیه گر محدودیت گرا عمدتا از پایگاه‌های اطلاعاتی از پیش تعریف شده استفاده می‌شود که حاوی قوانین صریح در مورد نحوه ارتباط نیازهای مشتری با ویژگی‌های مورد استفاده است[5].

سامانه‌های توصیه گر ترکیبی[2] : این سامانه ها بر اساس ترکیبی از تکنیک‌های ذکر شده فوق ساخته می‌شود. یک سامانه ترکیبی تلاش می‌کند که از مزایای یک سامانه برای رفع معایب سامانه دیگر استفاده کند[5].

سامانه‌های توصیه گر جمعیت شناسی[3]: در این نوع سامانه، اقلام مبتنی بر جمعیت شناسی توصیه می‌شود. فرض بر این است که باید توصیه‌های مختلف برای مدل‌های جمعیتی مختلف تولید شود. بسیاری از وب سایت‌ها به راحتی راه حل‌های شخصی سازی موثر بر اساس جمعیت شناسی را می‌پذیرند. به عنوان مثال، کاربران به وب سایت‌های خاصی بر اساس زبان یا کشورشان هدایت می‌شوند. یا پیشنهادات ممکن است با توجه به سن کاربر سفارشی شود[6].

سامانه‌های توصیه گر دانش محور[4]: در این سامانه ها، اقلام بر اساس دانش حوزه خاص و انطباق ویژگی‌های آیتم خاص با نیازهای کاربران توصیه می‌شود. در این سامانه، یک تابع شباهت مقدار نیاز کاربر (شرح مسئله) و تطابق توصیه‌ها (راه حل‌های مسئله) را تخمین می‌زند. در اینجا نمره تشابه می‌تواند به طور مستقیم به عنوان ابزار توصیه برای کاربر تفسیر شود. سامانه‌های مبتنی بر دانش در ابتدای پیاده سازی بهتر از سایر سامانه‌های توصیه گر عمل می‌کنند، اما اگر این سامانه‌ها با اجزای یادگیری مجهز نباشند، ممکن است نتایج مطلوبی در اثر گذشت زمان ارائه ندهند[7,8].

سامانه‌های توصیه گر جامعه گرا[5]: این نوع سامانه‌های توصیه گر اطلاعات مربوط به روابط اجتماعی کاربران و ترجیحات دوستان کاربر را به دست می‌آورد. توصیه‌ها بر اساس رتبه بندی هایی است که توسط دوستان کاربر ارائه شده است. در واقع این سامانه‌های توصیه گر در

1 Case-based systems
2 Hybrid recommender systems
3 Demographic
4 Knowledge-based systems
5 Community-based

پی افزایش شبکه های اجتماعی هستند و امکان دستیابی ساده و جامع از اطلاعات مربوط به روابط اجتماعی کاربران را فراهم می‌کنند[9] . در ادامه، مروری بر پژوهشهای انجام شده در این حوزه ارائه می‌گردد:

سامانه‌ی نشریاب الزویر[1] از جمله سامانه‌های شبیه به سامانه ژورنال یاب می‌باشد. در سامانه‌ی نشریاب الزویر از الگوریتم سامانه‌های توصیه‌گر استفاده می‌شود [10]. الگوریتم رتبه بندی توصیه گر نشریه به دو بخش تقسیم می‌شود. بخش اول مطابق با پرس و جو شده ارسال شده به مقالات موجود در پایگاه داده است. برای این منظور از الگوریتم Okapi BM25 استفاده شده است [11]. Okapi BM25 الگوریتمی است که به طور گسترده‌ای در زمینه بازیابی اطلاعات مورد استفاده قرار می‌گیرد. این الگوریتم، اسناد را مطابق با ارتباط آنها با یک پرس و جو، جستجو و رتبه بندی می‌نماید. به طور معمول، ورودی کیسه‌ای از کلمات[2] و خروجی مجموعه‌ای از اسناد با نمرات و رتبه بر اساس کلمات پرس و جو در هر سند، بدون در نظر گرفتن رابطه بین اصطلاحات پرس و جو در یک سند، می‌باشد. بخش دوم الگوریتم رتبه بندی توصیه گر نشریات، نمرات هر یک از مقالات را به نمرات مجلات منتقل می‌کند. این مرحله به زیر مراحل زیر تقسیم می‌شود[10]:

۱. یک میلیون مقاله با بالاترین نمره BM25 را از فهرست مقاله رتبه بندی نگه داشته و نشریه و حوزه‌های علمی نشریه را که هر مقاله به آن تعلق دارد را، پیدا شود.

۲. اگر کاربر نهایی قبلا یک دامنه را انتخاب کرده است، تمام اسنادی را که به این دامنه تعلق ندارند حذف می‌شود. اگر کاربر نهایی یک دامنه برای متن ورودی را انتخاب نکند، این مرحله حذف می‌گردد.

۳. محاسبه میانگین نمره BM25 در هر نشریه با استفاده از میانگین نمرات تمام مقالات منتشر شده در همان نشریه.

در زمینه سامانه‌های توصیه‌گر نشریه، در حال حاضر سامانه هایی وجود دارند که جستجوی مقالات مشابه را انجام می‌دهند[12]. به عنوان مثال، سامانه پاب‌مد[3] تابعی را برای

1 Elsevier Journal Finder
2 Bag-of-word
3 Pubmed

جستجوی رکوردهای مشابه از رکوردهای موجود در پایگاه اطلاعاتی مدلاین[1] ارائه می‌دهد[13] در سامانه ایتی بلاست[2] جستجو در چکیده مقالات را جهت توصیه نشریه انجام می‌شود[14]. سامانه ماندلی[3] مقالات مشابه را بر اساس مقالاتی که قبلا منتشر شده اند، جستجو کرده اما مجلات را توصیه نمی کنند[15].

در وبسایت آمازون[4] از الگوریتم‌های توصیه گر برای شخصی سازی فروشگاه آنلاین برای هر مشتری استفاده می‌شود[16]. پیشنهادات این فروشگاه آنلاین با توجه به علاقه مشتری تغییر می‌کند. برای مثال عناوین برنامه نویسی به یک مهندس نرم افزار و اسباب بازی‌های کودک را به یک مادر پیشنهاد داده می‌شود[17]. در آمازون از فیلترینگ مبتنی بر آیتم[5] استفاده می‌شود. بر خلاف فیلترینگ مشارکتی سنتی، محاسبات آنلاین این الگوریتم مستقل از تعداد مشتریان و تعداد اقلام موجود در کاتالوگ محصول است. این الگوریتم اقلام با کیفیت بالا را در زمان واقعی توصیه می‌کند . برای تعیین مشابه ترین آیتم برای یک مورد خاص، الگوریتم با ایجاد مواردی که بیشتر مشترکان تمایل به خرید آن کالا را دارند، یک جدول تشابه- کالا ساخته می‌شود. می‌توان ماتریس محصول-به-محصول[6] را با تکرار از طریق تمام جفت‌ها و ساخت یک متریک شباهت برای هر جفت ایجاد کرد. با این حال، بسیاری از جفت محصولات دارای کاربران مشترکی نیستند و بنابراین این روش در شرایط پردازش زمان و استفاده از حافظه ناکارآمد است. در الگوریتم تکرار پذیر زیر روش بهتری را با محاسبه شباهت بین یک محصول واحد و تمام محصولات مرتبط فراهم می‌کند:

```
For each item in product catalog, I1
  For each customer C who purchased I1
    For each item I2 purchased by customer C
      Record that a customer purchased I1 and I2
    For each item I2
      Compute the similarity between I1 and I2
```

1 Medline
2 eTBLAST
3 Mendeley
4 Amazon.com
5 item-based
6 product-to-product

می‌توان شباهت بین دو مورد را با روش‌های مختلف محاسبه کرد. اما یک روش معمول استفاده از اندازه گیری کسینوسی است، که در آن هر بردار مربوط به یک آیتم است، نه یک مشتری، و اندازه ابعاد بردار M مربوط به مشتریانی است که این آیتم را خریداری کرده اند. این محاسبات آفلاین از جدول تشابه-کالا بسیار زمانبر بوده و در بدترین حالت ($O(N^2M)$ است. در عمل، این زمان به $O(NM)$ نزدیک تر است، زیرا اکثر مشتریان خرید بسیار کمی دارند. نمونه برداری از مشتریانی که عناوین با بهترین فروش را خریداری می‌کنند زمان اجرا را با اندکی کاهش در کیفیت، کمتر می‌کند. با توجه به جدول تشابه- کالا، موارد مشابه هر خرید کاربر توسط این الگوریتم پیدا شده، آنرا رتبه بندی کرده و دسته‌ای از آیتم‌ها را جمع آوری نموده و سپس موارد محبوب ترین یا مرتبط‌ترین را توصیه می‌کند. این محاسبات بسیار سریع بوده و بر اساس تعداد مواردی است که کاربر خرید کرده و یا امتیاز داده است[18]. می‌و همکاران [19] یک سامانه توصیه گر ویدیویی آنلاین را به نام ویدیوریچ[1] پیشنهاد کردند که ویدیوها را مطابق با ویدئو فعلی بدون نیاز به پروفایل کاربر نمایش می‌دهد. ارتباط ویدیوها با ویژگی‌های متنی آنها (برچسب ها، کلمات کلیدی) تعیین می‌شود. فیلم‌ها با ویژگی‌های متنی مرتبط هستند. در پژوهش دیگر، کریشناکومار، سامانه توصیه گر ویدیویی آنلاین به نام ریکو[2] را ایجاد کرد. درسامانه ریکو، پروفایلی برای هر کاربر ساخته شده و داده‌های مورد علاقه کاربر به طور صریح از وی اخذ شده است. در این سامانه، اطلاعات جمع آوری شده از کاربر را با داده‌های مشابه جمع آوری شده برای سایر کاربران مقایسه کرده و یک لیست از اقلام توصیه شده را تهیه می‌کند. از پروفایل کاربری برای مطابقت اقلام با علاقه کاربر استفاده می‌گردد [20].

سان و کیم یک سامانه توصیه‌گر را برای یافتن مقالات علمی ارائه کردند. روش پیشنهادی آنها مبتنی بر شبکه چند سطحی استنادی[3] است و همه مقاله هایی که به طور غیر مستقیم مرتبط با مقاله مورد نظر هستند را در نظر گرفته و روابط ساختاری و معنایی بین آنها را بررسی کند. این سامانه توصیه‌گر مقالات مفیدی را که مربوط به موضوع تحقیق کاربر است را توصیه می‌نماید[21].

1 VideoReach
2 Recoo
3 Multileval Citation Network

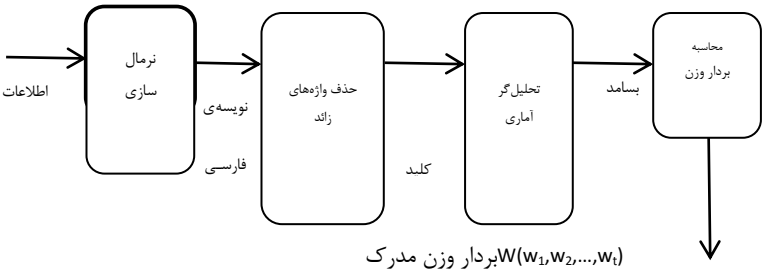
در این مقاله پیشینه تحقیات در بخش دوم و در بخش سوم معماری سامانه ژورنال‌یاب ارائه شده است. آزمون‌ها و نتایج در بخش چهارم شرح داده می‌شود و در انتها نتیجه گیری بیان می‌گردد.

۲-معماری سامانه توصیه‌گر ژورنال یاب

امروزه با گسترش بکارگیری رایانه، کاربرد سامانه‌های توصیه‌گر توسعه یافته است. همچنین با توجه به حجم و انبوه اطلاعاتی که در پیکره‌ها، بانک‌های اطلاعاتی و سطح وب ذخیره شده است، یافتن مدارک یا صفحات مرتبط و پیشنهاد دادن آن به کاربران می‌بایست هوشمندانه و دقیق باشد. هدف از این پژوهش طراحی سامانه‌ای برای پیشنهاد نشریه به پژوهشگران بر اساس اطلاعات کتابشناختی مقاله می‌باشد.

نشریات دارای حوزه‌های گوناگونی هستند، همچنین برخی از نشریات در حوزه‌های بین رشته‌ای فعالیت می‌کنند. بنابراین یافتن نشریه مناسب برای چاپ مقاله علمی می‌تواند یکی از چالش‌های پژوهشگران پس از نگارش متون علمی باشد. با توجه به اینکه پردازش زبان طبیعی برای متون زبان فارسی همواره با چالش‌هایی روبرو است، لذا سامانه توصیه‌گر باید راه حل‌هایی برای مشکلات خط فارسی در رایانه نیز لحاظ کند. سامانه توصیه‌گر پیشنهادی از روش ترکیبی تحلیل آماری[1] و مقایسه شباهت براساس شباهت مشارکتی استفاده می‌کند. در این روش متن اطلاعات کتابشناختی مقاله مورد نظر کاربر در ابتدا مورد پیش پردازش قرار می‌گیرد. ابتدا نویسه‌های غیرفارسی با معادل‌های آن در زبان فارسی جایگزین می‌شود. سپس واژه‌های زائد[2] از متن حذف می‌شوند. سپس تحلیل‌گر آماری واژه‌های کلیدی مستخرج را از نظر آماری پردازش می‌نماید و برای پرسش بسامد واژه‌های کلیدی را محاسبه می‌کند. بر اساس بسامدهای مستخرج، برداری از وزن واژه‌های کلیدی تشکیل می‌گردد. این بردار نماینده‌ی این اطلاعات کتابشناختی (مدرک) است. شکل ۱ عملکرد این ماژول را به تصویر می‌کشد.

1 Statistical Method
2 Stop words



شکل ۱- عملکرد سامانه توصیه گر پیشنهادی

برای محاسبه وزن هر کلیدواژه، از فرمول ۱ برای محاسبه‌ی وزن استفاده می‌شود. همانطور که در فرمول ۱ نشان داده شده است، برای محاسبه‌ی وزن کلید واژه بسامد نرمال شده‌ی واژه در بسامد معکوس مدرک آن واژه ضرب می‌شود تا وزن واژه بدست آید.

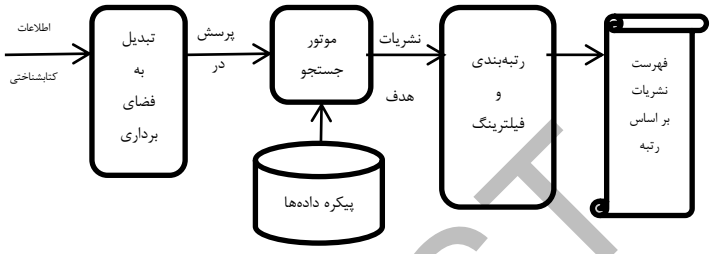
$$w_i = tf_i * log_2 df_i \quad (۱)$$

برای محاسبه‌ی tf_i بسامد واژه بر بیشینه بسامد واژه‌ای آن مدرک تقسیم می‌شود. در این صورت مقدار بسامد واژه‌ای به مقداری بین صفر تا یک نرمال می‌شود. بسامد معکوس مدرک نیز با شمارش تعداد مدارکی که شامل واژه‌ی i هستند محاسبه می‌شود.در این مرحله، هر مدرک با یک بردار مدل می‌شود که نشان دهنده‌ی آن مدرک در فضای برداری است.

پیاده‌سازی مدرک در فضای برداری در مدل پیشنهادی از ترکیب‌های یونی‌گرم[1] و بای‌گرم[2] واژه‌های اطلاعات کتابشناختی مدرک ساخته می‌شود. برای این منظور ترکیب‌های تکی و دو به دو از واژه‌های متن اطلاعات کتابشناختی تهیه می‌شود و بردار نماینده هر مدرک شامل کلیدواژه‌های تکی (یونی‌گرم) و ترکیب‌های دو به بدو (بای‌گرم) از واژه‌ها است. روش محاسبه‌ی وزن ترکیب‌های بای‌گرم نیز فرمول (۱) است. پس از محاسبه‌ی بردار مدرکی که به عنوان ورودی داده شده است، باید این بردار با بردارهای مستخرج از پیکره‌ی مقالات مقایسه شود تا بتوان مقالاتی شبیه به این حوزه را شناسایی نمود. اگر مقالات هدف دارای شباهت بالایی به

1 Uni-gram
2 Bi-gram

مقاله ورودی باشند، قاعدتاً نشریه‌هایی که این مقالات در آن‌ها چاپ شده‌اند می‌توانند به عنوان نشریات پیشنهادی توسط سامانه‌ی توصیه‌گر پیشنهاد شود. برای این منظور می‌بایست تمام مدارک موجود در پیکره نیز به صورت بردار مدل شوند. بنابراین در یک فرآیند برون‌خطی[1]، از تمامی مدارک پیکره اطلاعات کتابشناختی استخراج می‌شود و سپس با استفاده از ماژول شکل (۱) و محاسبه‌ی ترکیب‌های واژگانی تکی و دو به دو بردار هر مدرک تهیه و ذخیره می‌شود.



شکل۲- معماری سامانه توصیه‌گر

شکل (۲) معماری سامانه‌ی توصیه‌گر را نشان می‌دهد. ابتدا اطلاعات کتابشناختی مقاله مورد نظر کاربر (پرسش) از ورودی دریافت می‌شود. سپس این اطلاعات طبق الگوریتم بکارگرفته در شکل (۱) به برداری از وزن‌ها تبدیل می‌شود. اطلاعات پیکره‌ی مقالات قبلاً و بصورت برون‌خطی به فضای برداری منتقل شده است. این اطلاعات در این مرحله به عنوان بانک مقالات مورد استفاده قرار گرفته و با مقایسه برداری زاویه بین بردار پرسش و مقالات پیکره محاسبه می‌شود. با در نظر گرفتن یک آستانه برای شباهت شبیه‌ترین مقالات به پرسش استخراج شده و از طریق آن‌ها نشریات هدف شناسایی می‌شود. پس از شناسایی نشریات هدف می‌بایست فهرست نشریات براساس پرسش کاربر فیلتر شده و نشریات خارج از دامنه‌ی پرس‌وجو از مجموعه‌ی جواب‌ها حذف شود. در ضمن نشریات باقیمانده نیز بر اساس فاکتورهای چندگانه ازقبیل شباهت با مقاله‌ی کاربر، رتبه نشریه در سامانه‌ها رتبه‌بندی و ... رتبه‌بندی می‌شود و در نهایت جواب‌ها بصورت مرتّب براساس رتبه‌ی نشریه به کاربر نمایش داده می‌شود. در سامانه پیشنهادی برای محاسبه رتبه‌ی نهایی از فاکتورهای رتبه مقالات نشریه، رتبه نشریه

1 Off-line

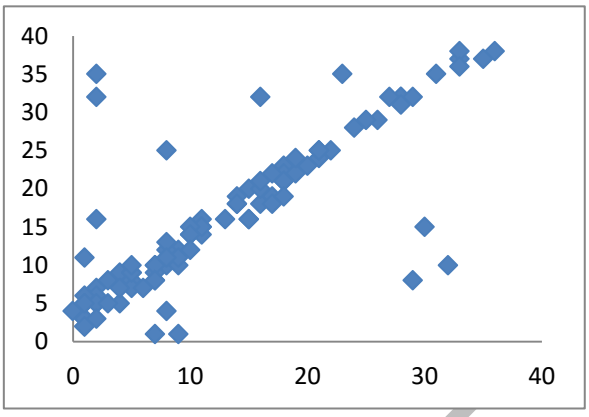
در نظام رتبه‌بندی و میزان مشابهت موضوعات نشریه با پرسش استفاده می‌شود. برای تهیه معیار واحد، از ترکیب خطی وزن‌دار این سه معیار مطابق با فرمول (۲) استفاده می‌شود.

$$Rank_J = w_1.R_i^j + w_2.IF_i^j + w_3.T_i^j \quad (۲)$$

در فرمول (۲)، R_i نماینده‌ی رتبه کسب شده توسط نشریه با معیار شباهت مقالاتش نسبت به پرسش است، IF^J نشان دهنده‌ی رتبه‌ی نشریه در نظام رتبه بندی نشریات و T^j نماینده‌ی معیار شباهت موضوعات نشریه با پرسش کاربر می‌باشد. این ترکیب خطی با استفاده از ضرایب w_i به صورتی تنظیم می‌شود تا نشریات مرتبط بهترین جواب را بدست آورند. برای این منظور در مجموعه‌ی آموزش ضرایب بصورتی تنظیم می‌شود که جواب نهایی در فهرست سه نشریه برتر بالای لیست مشاهده گردد.

۳-آزمون‌ها و نتایج

برای آزمودن سامانه‌ی توصیه‌گر نیاز است داده‌هایی برای آموزش و آزمایش سامانه گردآوری شود. بنابراین برای این منظور مجموعه داده‌های برچسب‌دار مورد نیاز است که مشخص کند برای هر مقاله نشریه هدف چیست.با توجه به طراحی و پیاده سازی پایگاه نشریات الکترونیکی متن کامل فارسی در مرکز منطقه‌ای از سال ۱۳۸۲ [22] بسترمناسب جهت پیاده سازی این آزمون فراهم است. در این راستا با استفاده از پایگاه اطلاعاتی مقالات فارسی مرکز منطقه‌ای اطلاع رسانی علوم و فنّاوری، مجموعه‌ای دارای حدود ۹ هزار عضو از مقالات علمی پژوهشی که دارای چکیده هستند، انتخاب شده است. برای هر مقاله در این مجموعه، نشریات هدف به صورت خودکار از پایگاه اطلاعاتی استخراج شده‌اند، نشریه‌ای که مقاله در آن چاپ شده است به عنوان مرتبط ترین جواب علامت خورده است. این مجموعه به دو بخش آموزش و آزمون شکسته می‌شود. بخش آموزش برای تنظیم پارامترهای رتبه‌بند سامانه مورد استفاده قرار گرفته و قسمت آزمون برای آزمایش دقت نهایی سامانه مورد استفاده قرار گرفته است. شکل (۳) نمودار ضریب همبستگی پیرسون را برای داده‌های آزمون و پاسخ‌هایی که سامانه برای مجموعه‌ی آزمون بدست آورده است نشان می‌دهد. همانطور که شکل (۳) نشان می‌دهد، سامانه‌ی توصیه‌گر با همبستگی بالا (۰,۸۱) با جواب‌های واقعی توانسته است پاسخ صحیح را توصیه کند.



شکل ۳- نمودار ضریب همبستگی پاسخ سامانه توصیه‌گر و جواب‌های واقعی (۴۰ پاسخ برتر)

جدول (۱) میانگین دقت سامانه‌ی توصیه‌گر را نشان می‌دهد. با مراجعه به داده‌های جدول (۱) مشخص است که سامانه‌ی توصیه‌گر با بکارگیری زیر سامانه رتبه‌بند، می‌تواند پاسخ صحیح را در رتبه‌ی درستی نشان دهد. میانگین رتبه‌ی پاسخ درست برای مجموعه‌ی آزمون ۱۰٫۵ است. به عبارت دیگر به طور متوسط در ۱۰ پاسخ اول پاسخ نشریه‌ی مرتبط درست نمایش داده شده است. رتبه‌ی پاسخ مهم است، زیرا اگر پاسخ درست در رتبه‌ی صحیح قرار نگیرد با بالا بردن آستانه فیلترینگ برای نمایش پاسخ، سطح نویز بالا رفته و می‌تواند موجب کاهش دقت سامانه شود. اگر آستانه بسیار سخت‌گیرانه انتخاب شود، ممکن است برخی از توصیه‌های مناسب برای پرسش از فهرست جواب خارج شود. همچنین باید توجه داشت که تعداد جواب‌هایی که می‌توان به کاربر ارائه داد محدود است.

جدول ۱: میانگین دقت سامانه‌ی توصیه‌گر

۴۰ پاسخ برتر	
۰٫۷۹	میانگین دقت (MAP)

رتبه‌ی قرارگیری پاسخ در فهرست جواب بسیار مهم است زیرا تعداد پاسخ‌ها براساس

آستانه محدود بوده و کاربر انتظار دارد که بالای فهرست جواب‌های دقیق‌تری را مشاهده نماید. در این سامانه‌ی توصیه‌گر کاربر در زمان پرسش دامنه‌ی جست‌وجو را محدود می‌کند(شکل ۴). این محدودیت می‌تواند برای حوزه‌ی نشریه (علوم انسانی، مهندسی، پزشکی و ...)، رتبه‌ی نشریه (دارای ضریب تاثیر، علمی پژوهشی و ...)، بسامد انتشار نشریه و پارامترهای دیگر تنظیم شود. این تنظیمات می‌تواند سامانه‌ی توصیه‌گر را در یافتن جواب صحیح هدایت نماید.

معیار دیگری که برای سنجش صحت فهرست پاسخ سامانه‌ی توصیه گر کاربرد دارد، معیار میانگین امتیاز متقابل[1] است (فرمول ۳)[23]. این معیار یک سنجش آماری است که برای هر پردازشی که فهرستی از جواب‌ها را برای یک پرسش تهیه می‌کند، قابل استفاده است.

شکل ۴: سامانه ژورنال یاب RICeST

امتیاز متقابل برای جواب‌های یک پرسش برابر با وارون ضربی رتبه‌ی اولین جواب درست است. میانگین امتیاز متقابل با بدست آوردن میانگین مقدار امتیاز متقابل برای تمام

1 Mean Reciprocal Rank (MRR)

پرسش‌هایی که در مجموعه‌ی Q قرار دارند بدست می‌آید. در فرمول (۳) مقدار $rank_i$، رتبه‌ی اولین جواب درست در مجموعه‌ی پاسخ‌های پرسش Q است.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{Rank_i} \quad (۳)$$

با استفاده از این معیار نیز آزمون انجام شده بروی سامانه‌ی توصیه‌گر ارزیابی می‌گردد. و نهایتا میانگین امتیاز متقابل برای سامانه‌ی توصیه‌گر نشریات (ژورنال یاب) برابر ۰,۵۳ بدست می‌آید.

۴-نتیجه‌گیری

نتایج بدست آمده نشان می‌دهد که الگوریتم ترکیبی بکارگرفته شده در سامانه‌ی ژورنال یاب کارا است و می‌تواند بخوبی نشریه هدف را شناسایی نماید. همچنین استفاده از ضریب همبستگی پیرسون نشان می‌دهد که پاسخ‌ها با جواب‌های واقعی هم‌راستا بوده و همبستگی خوبی بین پاسخ‌های سامانه و پاسخ‌های واقعی برقرار است. برای آنکه بتوان از سامانه بصورت عملی استفاده نمود، کاربر باید با محدود سازی دامنه‌ی جستجو، نویز را از مجموعه‌ی پاسخ‌ها حذف نماید که در این صورت دقت پاسخ قابل قبول خواهد بود. دقت پاسخ‌ها برای سامانه‌ی ژورنال‌یاب قابل قبول است. آزمون‌ها نشان داد که دقت بدست آمده در حوزه‌ی نشریات فارسی می‌تواند پاسخ‌های صحیح را در میانگین حدود رتبه‌ی ۱۰ شناسایی نماید. همچنین معیار میانگین امتیاز متقابل نیز نشان می‌دهد که رتبه‌ی پاسخ‌های صحیح در فهرست نتایج، امتیاز قابل قبولی است و میانگین امتیاز کسب شده برای رتبه‌ی پاسخ‌های صحیح قابل قبول است.با توجه به کارآیی سامانه، ژورنال‌یاب می‌تواند برای جستجو در نشریات و یافتن نشریه هدف مورد استفاده پژوهشگران قرار گیرد. برای توسعه‌ی سامانه در آینده، می‌توان نتایج رتبه‌ی مقالات شبیه به نشریه را با بکارگیری مجموعه منابع مقالات و نشریات و ساخت شبکه معنایی این داده‌ها ارتقاء داد. همچنین با استفاده از مدل مفهومی موضوعی و آنالیز رابطه‌ی پنهانی میان واژه‌ها و نمایه سازی معانی پنهان[1] می‌توان وزن‌دهی بردارها را بهبود بخشید.

1 Latent Semantic Indexing

منابع

[1] Nichols, D.M., 1997. Implicit ratings and filtering., *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*, Hungary, pp.31-36.

[2] Guo, Q. and Agichtein, E., 2012. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior, *proceedings of the 21st international conference on World Wide Web*, pp. 569–578.

[3] Zhu, Y., He, L., and Wang, X., 2012.User interest modeling and self-adaptive update using relevance feedback technology. *Procedia Engineering*, 29,pp. 721–725.

[4] Park, Y.J., 2013. An adaptive match-making system reflecting the explicit and implicit preferences of users. *Expert Systems with Applications*, 40(4), pp.1196–1204.

[5] Neumann, A.W.2009. *Recommender Systems for Information Providers*. Physica-Verlag,A Springer Company.

[6] Kelly, D. and Belkin, N. J., 2004. Display time as implicit feedback: understanding task effects. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '04*, New York, NY, USA, pp. 377–384.

[7] Oard, D. W. and Kim, J.,2001. Modeling information content using observable behavior. *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*, USA, pp.38-45.

[8] Kelly, D. and Teevan, J., 2003.Implicit feedback for inferring user preference: a bibliography,*ACM SIGIR Forum, 37*,pp 18-28.

[9] Guo , Q. and Agichtein, E.,2012. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. Proceedings *the 21st international conference on World Wide Web.* pp. 569–578.

[10] Buscher, G., White, R. W., Dumais, S. and Huang, J., 2012. Large-scale analysis of individual and task differences in search result page examination strategies. in *Proceedings of the fifth ACM international conference on web search and data mining*, pp. 373–383.

[11] Tyler, S. K., Wang, J. and Zhang, Y. , 2010. Utilizing re-finding for personalized information retrieval. *Proceedings of the 19th ACM international information and knowledge management*, pp. 1469–1472.

[12] Precision and recall, Retrieved from https://en.wikipedia.org/wiki/Precision_and_recall

[13] Kantrowitz, M.. 2000. Stemming and its effects on TFIDF Ranking. *Proceedings of the 33st annual international ACM SIGIR conference on Research and development in information retrieval*, pp.357–359.

[14] Errami, M. 2007. ETBLAST: A web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Research*, 35(S2).

[15] Reiswig, J. 2010. Mendeley. *Journal of the Medical Library Association*. 98, (2010), pp.193–194.

[16]George, T. 2005. A scalable collaborative filtering framework based on co-clustering. *Fifth IEEE International Conference on Data Mining*, pp. 625–628

[17]Kay, J. 2006. Scrutable adaptation: Because we can and must. *Adaptive Hypermedia and Adaptive Web-Based Systems, 4th International Conference*, Dublin, Ireland, pp. 11–19.

[18]Linden, G., Smith, B., York, J.2003, Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing* ,7(1), pp.76–80.

[19]T. Mei, B. Yang, X. Hua, L. Yang, S. Yang, and S. Li, 2007.VideoReach: an online video recommendation system. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM*. pp. 767–768.

[20]A. Krishnakumar. 2007.Recoo: A Recommendation System for Youtube RSS Feeds. University of California, Santa Cruz, Tech. Rep.

[21] Son, J., & Kim, S. B. 2018. Academic paper recommender system using multilevel simultaneous citation networks. *Decision Support Systems*, *105*,pp. 24-33.

[۲۲] مهراد،جعفر، کلینی، سارا،۱۳۸۳."پیاده سازی پایگاه نشریات الکترونیکی متن کامل فارسی در کتابخانه منطقه‌ای علوم و تکنولوژی شیراز"، کتابداری و اطلاع رسانی، ۸۳-۴،۹۶.

[23] Radev, D. R. , Qi, H., Wu, H., Fan, W.2002. Evaluating web-based question answering systems. *Proceedings of LREC.*

خوشه‌بندی رباعیات عمر خیام با روش کا-میانگین

پروانه خسروی‌زاده* و **محمد رجب‌پور****

چکیده

در این پژوهش با پیاده‌سازی الگوریتم کا-میانگین، رباعیات عمر خیام (نسخه‌ی محمدعلی فروغی) خوشه بندی شد تا رایانه رباعیاتی که از دیدگاه واژگانی ناهمگون‌اند و آنها که همسان‌اند را آشکار سازد. ویژگی‌های به کار رفته برای خوشه‌بندی از یک سو «فراوانی واژگان» و از دیگر سو «فراوانی واژگان در وارون فراوانی سندها» بود. خوشه‌بندی هم با زدایش ایست‌واژه‌ها و هم با نگهداشت آنها انجام گرفت. فرایند خوشه‌بندی با شمار خوشه های گوناگون، از یک تا پنجاه خوشه، بارها از سر گرفته و یافته‌های عددی با یکدیگر سنجیده شد. بدین سان با در نگر آوردن دو ویژگی بالا، بود یا نبود ایست‌واژه‌ها و تعداد خوشه‌ها، رباعیات به ۲۰۰ شیوه‌ی گوناگون خوشه‌بندی شدند. یافته‌های این پژوهش می‌تواند در بازشناسی اصالت رباعیات خیام روشنگر باشد و به منتقدان ادبی یاری رساند.

واژه‌های کلیدی: رباعیات عمر خیام، خوشه‌بندی، الگوریتم کا-میانگین، اصالت متن، زبان‌شناسی رایانشی.

۱. مقدمه

بهره‌گیری از شیوه‌های آماری در بازشناسی الگو از دهه‌های پایانی قرن بیستم آغاز شد. در این زمینه هارتیگان [۱] را می‌توان از پیشگامان شناساندن الگوریتم خوشه‌بندی کا-میانگین دانست. هارتیگان و دیگران [۲]، این شیوه‌ی خوشه‌بندی را شرح می‌دهند و به توصیف مراحل انجام کار می‌پردازند. فوکوناگا [۳] در راستای معرفی روش‌های آماری در بازشناخت الگو، به تفکیک تکنیک‌هایی که در خوشه‌بندی به کار می‌روند می‌پردازند. راسموسن [۴] نیز خوشه‌بندی را توصیف می‌کند و روش‌های مختلف آن را برمی‌شمارد. آلن [۵] و مانینگ [۶] از

* استادیار گروه زبان‌شناسی رایانشی، دانشگاه صنعتی شریف، khosravizadeh@sharif.ir
** کارشناس ارشد زبان‌شناسی رایانشی، دانشگاه صنعتی شریف، mhmd.rjbpr@gmail.com

خوشه‌بندی و الگوریتم کا-میانگین در پردازش زبان طبیعی بهره می‌گیرند و ویژگی‌های آن را شرح می‌دهند. چنین برمی‌آید که خوشه‌بندی بیشتر هنگامی به کار می‌رود که با دادگان برچسب‌زده نشده‌ی انبوهی سر و کار داریم. در اینجا تلاش می‌شود از خوشه‌بندی کا-میانگین برای بازشناسی اصالت رباعیات عمر خیام استفاده شود.

از الگوریتم خوشه‌بندی کا-میانگین می‌توان در تشخیص اصالت یک متن و تفکیک بخش‌های اصیل از افزوده‌های متن به‌ویژه در مورد آثاری که در رهگذر زمان ثبت شده‌اند و در اصالت بخشهایی از اثر تردید وجود دارد بهره برد. از این رو، در پژوهش حاضر تلاش شده است تا با پیاده‌سازی الگوریتم کا-میانگین رباعیات حکیم عمر خیام خوشه‌بندی شده و آن دسته از رباعیاتی که از نظر واژگانی ناهمگون هستند از دیگر رباعیات خیام تفکیک گردند.

نمونه‌ی شناخته شده از نسخه‌ای که پژوهشگران در آن رباعیات را دسته‌بندی کرده‌اند نسخه‌ای است که هدایت [۷] منتشر کرده است. هدایت رباعیات خیام را بر اساس موضوعات مطرح شده در آنها دسته‌بندی کرده و رباعیاتی را که با جهان‌بینی شاعر ناسازگارند کنار می‌نهد. اما به نظر می‌رسد نسخه‌ی محمدعلی فروغی [۸] و [۹] بیشترین اقبال را در میان ناشران در سال‌های اخیر داشته است و می‌توان آن را از معتبرترین نسخه‌ها دانست. در اینجا یک چاپ پیش از انقلاب [۸] و یک چاپ پس از انقلاب [۹] این نسخه برگزیده شده است و دادگان پردازش شده از دو تارنمای یاد شده در منابع [۱۰] و [۱۱] گرفته شده‌اند. دلیل انتخاب یک نسخه‌ی پیش از انقلاب و یک نسخه پس از انقلاب در نظر گرفتن تغییرات احتمالی در املای کلمات و امکان وجود دگرگونی‌هایی در رسم‌الخط بوده است. انتخاب این دو نسخه با احتمال وجود تفاوتهای ویرایشی و یا ممیزی صورت گرفته به ارتقا سطح به‌هنجارسازی منجر شده است. بن‌مایه‌ی اصلی دادگان پژوهش نسخه‌ی اینترنتی رباعیات موجود در فضای مجازی بود. اما در مرحله‌ی به‌هنجارسازی و ویرایش، دادگان با دو نسخه‌ی قدیم و جدید مطابقت داده شدند.

۲. روش انجام پژوهش

در روش خوشه‌بندی کا-میانگین در هر خوشه‌بندی N نمونه D بعدی داریم: {x1, x2, . . . , xN}. در اینجا هر نمونه یک رباعی خیام است و N، برابر با ۱۷۸ است که

همان تعداد رباعیات گنجانده شده در نسخه‌ی محمدعلی فروغی است. برای این که دقت خوشه‌بندی افزایش یابد همه‌ی رباعیات به شیوه‌ی دستی به‌هنجارسازی شدند. هر رباعی با بردار ویژگی وابسته (TF یا TF-IDF) شناسانده شد. انگیزه‌ی انجام این پژوهش، خوشه‌بندی رباعیات در K خوشه با مرکزهای $\{\mu_1, \mu_2, \ldots, \mu_K\}$ بود. در آغاز، مرکزهای خوشه‌ها به شیوه‌ی تصادفی مقداردهی شدند. برای این که یافته‌های پژوهش از دیدگاه علمی تکرارپذیر باشند مقدار اولیه‌ی هر کدام از مرکزها برابر با مختصات بردار دارای اندیس حاصلضرب شماره‌ی آن مرکز در حاصل‌تقسیم جزءصحیح عدد ۱۷۸ بر k (تعداد خوشه‌ها) انگاشته شد.

$$\mu_n = X_{(n \times ([178 \div k])} \quad (۱)$$

معادله ۱: مقدار اولیه‌ی بردار مرکز n اُم

سپس تا زمانی که مقدار مرکز خوشه‌ها تغییر می‌کرد، گام‌های زیر پیوسته از سر گرفته می‌شدند:

- بسته به کمیت‌های مراکز خوشه‌ها، نزدیک‌ترین خوشه به هر رباعی یافته می‌شد.
- با توجه به داده‌های هر خوشه، مختصات مرکز خوشه دوباره محاسبه می‌گردید.

برای سنجش درستی و دقت خوشه‌بندی، در هر حالت برای هر خوشه میانگین معیارهای انسجام درونی[1]، تفکیک برونی[2] و ضریب سیلوئت[3] محاسبه گردید و میانگین ضریب سیلوئت برای همه داده‌ها نیز نمایانده شد.

$$Cohesion(C_k) = \sum_{x \,\&\, y \in C_k} Similarity\ (x \,\&\, y) \quad (۲)$$

معادله ۲: محاسبه‌ی انسجام درونی

$$S\ eparation(C_i \& C_j) = \sum_{x \in C_i \& y \in C_k} Similarity\ (x \,\&\, y) \quad (۳)$$

معادله ۳: محاسبه‌ی تفکیک برونی

هنگامی که فاصله‌ی اقلیدسی یا به عبارتی دیگر فاصله‌ی متریک داریم، ضریب سیلوئت

1 Cohesion
2 Separation
3 Silhouette Coefficient

برای هر داده x_i از رابطه‌ی زیر به دست می‌آید:

$$S_i = \frac{b_i - a_i}{Max(b_i \& a_i)} \quad (۴)$$

معادله ۴: محاسبه‌ی میانگین ضریب سیلوئت در هنگام وجود فاصله‌ی اقلیدسی

در رابطه‌ی فوق a_i برابر است با فاصله داده x_i از تمام داده‌های دیگر در خوشه‌ی خودش که همان تعریف انسجام درونی است. برای به دست آوردن b_i نخست میانگین فاصله‌ی داده x_i از تمام داده‌های دیگر در K-1 خوشه‌ی دیگر محاسبه می‌شود که همان تفکیک برونی است و سپس کمترین مقدار به دست آمده به عنوان مقدار b_i انتخاب می‌گردد.

از آنجا که برای خوشه‌بندی متن‌ها، فاصله‌ی اقلیدسی چندان مناسب نیست و باید از فاصله‌ی کسینوسی بهره جست، در معادله محاسبه ضریب سیلوئت باید تغییراتی ایجاد می‌شد. نخست برای انتخاب b_i به جای کمترین کمیت، بیشترین مقدار برگزیده شد و سپس در صورت کسر جای a_i و b_i عوض گردید.

برای تک‌تک رباعیات ضریب سیلوئت محاسبه گردید و برای هر خوشه و تمام خوشه‌ها نیز میانگین ضریب سیلوئت به دست آمد.

بر حسب ویژگی «فراوانی واژگان» و با «نگهداشت ایست‌واژه‌ها» برای خوشه‌بندی از ۲ تا ۱۰ خوشه، ضریب سیلوئت به طور میانگین با تقریب سه رقم اعشار ۰,۲۰۸ است. برحسب ویژگی «فراوانی واژگان» و با «زدایش ایست‌واژه‌ها» برای خوشه‌بندی از ۲ تا ۱۰ خوشه، ضریب سیلوئت به طور میانگین با تقریب سه رقم اعشار ۰,۳۹۷ است. برحسب ویژگی «فراوانی واژگان در وارون فراوانی سندها» و با «نگهداشت ایست‌واژه‌ها» برای خوشه‌بندی از ۲ تا ۱۰ خوشه، ضریب سیلوئت به طور میانگین با تقریب سه رقم اعشار ۰,۲۳۹ است. برحسب ویژگی «فراوانی واژگان در وارون فراوانی سندها» و با «زدایش ایست‌واژه‌ها» برای خوشه‌بندی از ۲ تا ۱۰ خوشه، ضریب سیلوئت به طور میانگین با تقریب سه رقم اعشار ۰,۳۵۳ است.

جدول ۱: میانگین ضریب سیلوئت برای خوشه‌بندی ۲ تا ۱۰ تایی

	فراوانی واژگان	فروانی واژگان در وارون فراوانی سندها
نگه‌داشت ایست‌واژه‌ها	۰,۲۰۸	۰,۲۳۹
زدودن ایست‌واژه‌ها	۰,۳۹۷	۰,۳۵۳

جدول ۲: میانگین ضریب سیلوئت برای خوشه‌بندی ۲ تا ۲۱ تایی

	فروانی واژگان در وارون فراوانی سندها
نگه‌داشت ایست‌واژه‌ها	۰,۳۱۷
زدودن ایست‌واژه‌ها	۰,۳۹۳

برحسب ویژگی «فراوانی واژگان در وارون فراوانی سندها» و با «نگهداشت ایست‌واژه‌ها» برای خوشه‌بندی از ۲ تا ۲۱ خوشه، ضریب سیلوئت به طور میانگین با تقریب سه رقم اعشار ۰,۳۱۷ است. برحسب ویژگی «فراوانی واژگان در وارون فراوانی سندها» و با «زدایش ایست‌واژه‌ها» برای خوشه‌بندی از ۲ تا ۲۱ خوشه، ضریب سیلوئت به طور میانگین با تقریب سه رقم اعشار ۰,۳۹۳ است.

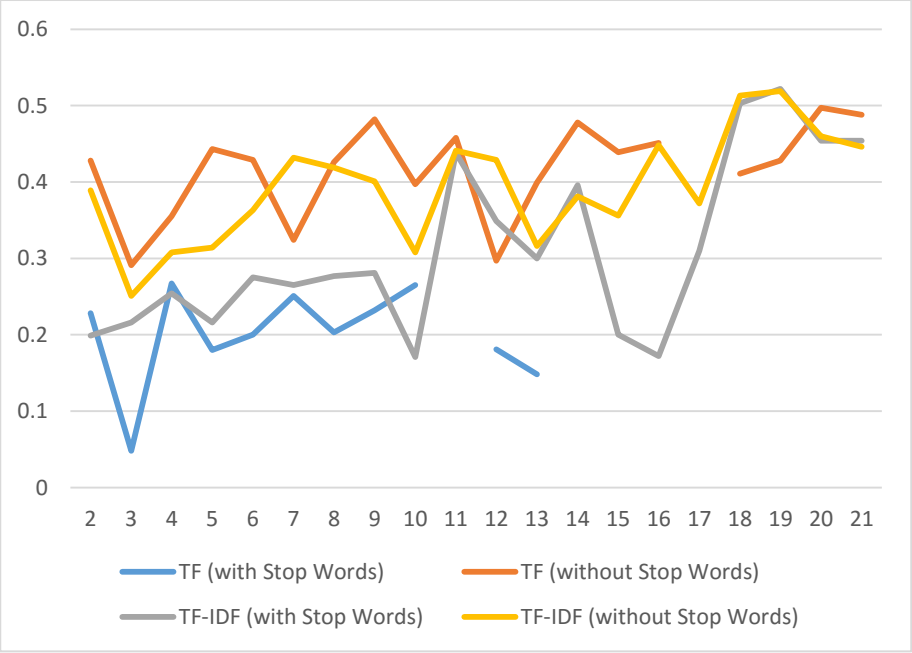
ضریب سیلوئت برای خوشه‌بندی ۲ تا ۲۱ تایی بر حسب ویژگی «فراوانی واژگان» از این رو محاسبه نشده است که در این صورت با «نگهداشت ایست‌واژه‌ها» در خوشه‌بندی ۱۱ تایی و خوشه‌بندی ۱۴ تایی به بالا دست‌کم یک خوشه‌ی تک‌عضوی وجود دارد که طبیعتاً انسجام درونی و ضریب سیلوئت برای آن تعریف‌پذیر نیست. در صورت «زدایش ایست‌واژه‌ها» نیز در خوشه‌بندی ۱۷ تایی یک خوشه تک‌عضوی وجود خواهد داشت.

زدایش ایست‌واژه‌ها موجب افزایش میانگین ضریب سیلوئت می‌شود. این افزایش هنگامی که ویژگی «فراوانی واژگان» به کار می‌رود برای خوشه‌بندی ۲ تا ۱۰ تایی حدود ۱,۹۱ برابر و هنگامی که ویژگی «فراوانی واژگان در وارون فراوانی سندها» مورد استفاده قرار می‌گیرد حدود ۱,۴۸ برابر است. به عبارت دیگر، در حالت نخست میانگین ضریب سیلوئت ۹۱ درصد افزایش می‌یابد و در حالت دوم این افزایش ۴۸ درصد است. اگر ملاک خوشه‌بندی ۲ تا ۲۱ تایی باشد، در حالت دوم تنها افزایشی ۲۴ درصدی رخ می‌دهد.

جدول ۳: مقدار کمینه و بیشینه‌ی ضریب سیلوئت

	فراوانی واژگان	فروانی واژگان در وارون فراوانی سندها
نگه‌داشت ایست‌واژه‌ها	$0.048 \leq S \leq 0.267$	$0.171 \leq S \leq 0.522$
زدودن ایست‌واژه‌ها	$0.291 \leq S \leq 0.497$	$0.251 \leq S \leq 0.519$

در خوشه‌بندی‌هایی که در هر خوشه دست‌کم دو رباعی قرار گرفته است بیشترین میانگین ضریب سیلوئت در تقسیم رباعیات به ۱۹ خوشه بر حسب «فراوانی واژگان در وارون فراوانی سندها» و با «نگهداشت ایست‌واژه‌ها» به دست می‌آید که با تقریب سه رقم اعشار برابر با ۰,۵۲۲ است و کمترین مقدار در تقسیم رباعیات به ۳ خوشه بر حسب «فراوانی واژگان» و با «نگهداشت ایست‌واژه‌ها» حاصل می‌شود که با تقریب سه رقم اعشار برابر با ۰,۰۴۸ است. در خوشه‌بندی بر حسب «فراوانی واژگان در وارون فراوانی سندها» و با «زدایش ایست‌واژه‌ها» نیز بیشترین مقدار میانگین ضریب سیلوئت در خوشه‌بندی ۱۹ تایی رخ می‌دهد.



شکل ۱: ضریب سیلوئت بر حسب تعداد خوشه‌ها

اگر بخواهیم با استفاده از روش خوشه‌بندی کا-میانگین رباعیاتی را که از لحاظ ویژگی‌های صوری زبانی نامرتبط و پرت هستند بیابیم، خوشه‌بندی با ویژگی «فراوانی واژگان» و با «نگهداشت ایست‌واژه‌ها» می‌تواند رهیافت بهتری برای مسأله باشد. چون بسیار زودتر از خوشه‌بندی‌های دیگر به حالتی می‌رسیم که دست‌کم یک خوشه دارای تنها یک عضو است. رباعی موجود در هر خوشه‌ی تک‌عضوی اگر دارای میانگین تفکیک بالایی نسبت به دیگر رباعیات در سایر خوشه‌ها باشد، احتمالاً دارای اصالت نیست و می‌بایست احتمال سراییده شدن آن توسط عمر خیام را بسیار کم دانست.

۳. نتایج

رباعیات زیر از جمله شعرهایی هستند که در هنگام خوشه‌بندی در خوشه‌های تک‌عضوی شکار شده‌اند و میانگین تفکیک پایینی نیز نسبت به رباعی‌های سایر خوشه‌ها داشته‌اند:

یک نان به دو روز اگر بُوَد حاصل مرد — ازکوزه شکسته‌ای دمی آبی سرد
ماموركم از خودی چرا باید بود — یا خدمت چون خودی چرا باید کرد
بر چرخ فلک هیچ کسی چیر نشد — وز خوردن آدمی زمین سیر نشد
مغرور بدانی که نخورده است تو را — تعجیل مکن هم بخورد دیر نشد
ای پیر خردمند پگه‌تر برخیز — وآن کودک خاکبیز را بنگر تیز
پندش ده گو که نرم نرمک می‌بیز — مغز سر کیقباد و چشم پرویز
هر صبح که روی لاله شبنم گیرد — بالای بنفشه در چمن خم گیرد
انصاف مرا ز غنچه خوش می‌آید — کو دامن خویشتن فراهم گیرد
وقت سحر است خیزای طرفه پسر — پر باده لعل کن بلورین ساغر
کاین یک دم عاریت دراین گنج فنا — بسیار بجوئی و نیابی دیگر
چندان که نگاه می‌کنم هر سویی — در باغ روان است ز کوثر جویی
صحرا چو بهشت است ز کوثرکم گوی — بنشین به بهشت با بهشتی رویی

بدیهی است که هر چه تعداد خوشه‌ها بیشتر شود، این روش برای شکار رباعیات نامرتبط ناکارآمدتر می‌گردد، زیرا احتمال قرار گرفتن رباعی‌هایی که به سایر رباعی‌ها شباهت دارند در یک خوشه به صورت منفرد بیشتر می‌شود.

از خوشه‌بندی رباعیات عمر خیام، می‌توان برای یافتن رباعیاتی که احتمال اصالتشان بیشتر است نیز استفاده کرد. می‌توان از خوشه‌بندی‌هایی که میانگین ضریب سیلوئتشان نزدیک ۰٫۵ است بهره برد و استدلال کرد که رباعیاتی که در یک خوشه با با میانگین ضریب سیلوئت بالا قرار دارند بسیار به هم شبیه‌اند و احتمالاً توسط خیام یا یک شاعر منحصر به فرد سروده شده‌اند.

علاوه بر یافتن رباعیات اصیل و غیراصیل، می‌توان از خوشه‌بندی برای تقسیم‌بندی معنایی شعرها نیز بهره جست. برای این منظور استفاده از ویژگی «فراوانی واژگان در وارون فراوانی سندها» چه با نگهداشت و چه با زدایش ایست‌واژه‌ها و روش «فراوانی واژگان» با «زدایش ایست‌واژه‌ها» مناسب‌تر است. در اینجا نیز باید دنبال خوشه‌بندی‌هایی گشت که میانگین ضریب سیلوئت بالایی دارند. برای مثال، اگر بخواهیم رباعیات را از دیدگاه معنایی به دو گروه تقسیم کنیم مشاهده می‌شود که استفاده از ویژگی «فراوانی واژگان» با «زدایش ایست‌واژه‌ها» بیشترین میانگین ضریب سیلوئت را با مقدار حدوداً ۰٫۴۲۸ به دست می‌دهد. بررسی نتایج این نوع خوشه‌بندی نشان می‌دهد که یک خوشه‌ی ۷۴ تایی و یک خوشه‌ی ۱۰۴ تایی به ترتیب با میانگین ضریب سیلوئت ۰٫۷۸۶ و ۰٫۰۷۱ وجود دارد. بیشتر رباعیاتی که دعوت به «خوش‌باشی» و «میگساری» و «لذت‌جویی» دارند در این خوشه واقع شده‌اند و رباعیات گوناگون دیگری که بیشتر با اندوه فلسفی آمیخته‌اند در خوشه‌ی دیگر قرار دارند.

۴. نتیجه‌گیری

بیشتر نسخه‌های رباعیاتی که تاکنون از عمر خیام منتشر شده‌اند دارای ترتیب الفبایی بر اساس حرف پایانی قافیه‌ی هر مصرع رباعی است. یافته‌های این پژوهش این امکان را به وجود می‌آورند که رباعیات به تفکیک معنایی به طور خودکار توسط رایانه چیدمان شوند و یا به طور نیمه‌خودکار با همکاری رایانه و ویراستار انسانی بخش‌بندی شوند.

پیشنهاد می‌شود تا در پژوهش‌های بعدی، رباعیات عمر خیام برچسب‌زنی نحوی نیز شود و ویژگی «الگوهای نحوی» با ویژگی «فراوانی واژگان» در کنار یکدیگر مورد استفاده قرار گیرد تا دقت خوشه‌بندی‌ها افزایش یابد. هم‌چنین می‌توان رباعی‌هایی که در نسخه‌ی محمدعلی فروغی وجود ندارند و در نسخه‌های دیگر موجودند را به مجموعه‌های رباعیات افزود تا اصالت

آنها نیز سنجیده شود.

منابع

[۱] Hartigan, J. A., 1975. *Clustering algorithms*, John Wiley& Sons, Inc.

[۲] Hartigan, J. A., and M. A. Wong, 1979. "Algorithm AS 136: A K-Means Clustering Algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 1979, pp. 100–108.

[۳] Fukunaga, K., 1990. *Introduction to Statistical Pattern Recognition.* Academic Press, San Diego, CA.

[۴] Rasmussen, E. 1992. "Clustering Algorithms", In *Information Retrieval Data Structures and*
Algorithms, W.B. Frakes and R. Baeza- Yates Eds., Prentice Hall, N J. USA, Chapter 16, pp. 548-573.

[۵] Allen, J., 1995, *Natural Language Understanding*, The Benjamin/Cummings Publishing Company Inc., USA.

[۶] Manning, C. D. and Schütze, H. 1999.
Foundations of Statistical Natural Language Processing. MIT Press, MA. USA.

[۷] هدایت، صادق، ۱۳۰۲. ترانه‌های خیام. تهران: انتشارات امیرکبیر.

[۸] فروغی، محمدعلی، ۱۳۵۴. *رباعیات عمر خیام*. تهران: انتشارات امیرکبیر.

[۹] فروغی، محمدعلی، ۱۳۸۹. رباعیات خیام. تهران: انتشارات سخن عشق.

[۱۰] ----------رباعیات خیام، ۱۳۸۵، ویکی‌نبشته، دریافت شده در آذرماه ۱۳۹۴ از؛
https://fa.wikisource.org/wiki/%D8%B1%D8%A8%D8%A7%D8%B9%DB%8C%D8%A7%D8%AA_%D8%AE%DB%8C%D8%A7%D9%85

[۱۱] ---------رباعیات عمرخیام، کتابخانه دیجیتال ری‌را، دریافت شده در آذرماه ۱۳۹۴ از؛
http://rira.ir/?page=view&mod=classicpoems&obj=poet&id=7

راهکارهایی برای ترجمه‌ی ماشینی از انگلیسی به فارسی از منظر ترتیب‌خطی

احمدرضا شریفی پورشیرازی* و محمد خانی**

چکیده

در این مقاله به منظور بهبود عملکرد سامانه‌ی ترجمه‌ی ماشینی، سامانه‌ی تغییر ترتیب خطی که سامانه‌ی مرتب‌سازی مجدد نامیده می‌شود، ارائه می‌گردد. اطلاعات مورد نیاز برای این سامانه از تحلیل‌گر استنفورد به‌دست آمده است و برای تبدیل جملات زبان انگلیسی به فارسی و به‌عبارتی، تبدیل یک زبان هسته‌ابتدا به هسته‌انتها، بر اساس ویژگی‌های مطروحه از سوی هایدر {۵} و تحلیل صورت گرفته در تحلیل‌گر استنفورد، قواعد و اصولی استخراج و به‌کار بسته می‌شوند. سپس این قواعد در قالب یک پردازشگر، به زبان انگلیسی اعمال می‌گردند تا ترتیب خطی نزدیک به زبان مقصد که فارسی است، به دست آید. رویکرد مطروحه، شامل یک مرحله‌ی پیش‌پردازشی است که در آن سامانه‌ی مرتب‌سازی مجدد برای شناسایی ویژگی‌های زبان‌های دارای ترتیب‌های خطی متفاوت، آماده‌سازی شده و تعلیم دیده است. بر همین اساس، ترتیب خطی جملات زبان مبدا قبل از ترجمه بر اساس آن اطلاعات زبانی زبان مقصد اصلاح می‌شوند که در سامانه‌ی مرتب‌سازی مجدد گنجانده شده است. مدل پیشنهادی روش موثری را برای تغییر ترتیب‌خطی زبان مبدا بر طبق ویژگی‌های زبان مقصد ارائه می‌کند. با این بررسی می‌توان نشان داد که استفاده از دانش زبان‌شناسی در پردازش داده‌های مورد بررسی، می‌تواند به پیشرفت‌های قابل توجهی در عملکرد ترجمه منتهی گردد.

واژه‌های کلیدی: ترجمه‌ی ماشینی، مرتب‌سازی مجدد، ترتیب خطی، زبان مبدا، زبان مقصد

۱. مقدمه

ترتیب خطی به این معناست که چه عنصری اول بیاید، چه عنصری در جایگاه بعد ظاهر شود، همچنین اینکه چه عنصری تلفظ شود یا نشود و ازآنجا که انسان صرفاً می‌تواند زنجیره‌ای

* دانشجوی دکتری زبانشناسی همگانی- دانشگاه شیراز- نویسنده‌ی مسئول- asharifipur@gmail.com.

** کارشناس ارشد زبانشناسی همگانی- مرکز منطقه ای اطلاع رسانی علوم و فناوری- khani7m@gmail.com

از لغاتی را به‌زبان بیاورد که دارای ترتیب خطی باشند، بنابراین تا عناصر موجود درون جمله دارای ترتیب خطی نشوند، قابلیت به بیان آورده شدن را نخواهند داشت. در زبان‌شناسی معمولاً ترتیب‌خطی پایه را با استفاده از یک فعل و موضوعات آن که شامل فاعل و مفعول می‌شود، تعریف می‌کنند. ازآنجا که زبان‌ها با توجه به ترتیب خطی لغاتی که در تولید جملات با معانی یکسان به‌کار می‌برند و به‌عبارتی، هسته‌ابتدا بودن یا هسته‌انتها بودنشان با یکدیگر تفاوت دارند، رعایت ترتیب خطی در ترجمه از زبان مبدا به زبان مقصد بسیار حائز اهمیت است. از همین‌رو، ترتیب خطی متون زبان مبدا می‌بایست در ترجمه به نحوی تغییر کنند که با حفظ معنای زبان مبدا، ترتیب خطی خوش‌ساختی را در زبان مقصد به‌دست دهند. بر همین اساس می‌توان گفت که عملیات تغییر ترتیب خطی نقش مهمی را در بهبود عملکرد نظام ترجمه‌ی ماشینی ایفا می‌کند. نظام‌های ترجمه‌ی ماشینی عبارت‌مبنا۱ تنها می‌توانند گستره‌ی محدودی از تغییر ترتیب لغات را با توسل به فهرستی از عبارت۲ به‌دست دهند و حتی به‌دست آوردن چنین تغییراتی در ترتیب لغوی نیز به‌علت میزان محدود داده‌های موجود، محدود می‌شود. به‌عنوان مثال، اگر در زبان مبدا صفت‌ها قبل از اسامی قرار بگیرند (همچون انگلیسی) و در زبان مقصد بعد از اسم بیایند (همچون فارسی) ما همچنان نیاز داریم تا در پیکره‌های مشابه توالی خاصی از صفت و اسم را بیابیم تا بتوانیم با توسل به فهرستی از عبارات، تغییر در ترتیب لغات را در کنترل بگیریم. نظام‌های عبارت‌مبنا، برای تولید ترتیب نزدیک به زبان مقصد بر مدل زبانی مقصد تکیه می‌کنند. این مسئله برطبق نظریات الاونیزان و پاپینی {۱} ناکارآمد است و این ناکارآمدی تلاش‌های بسیاری را برای غلبه بر مشکل ترتیب لغوی متفاوت در زبان‌ها موجب شده است.

یامادا و نایت {۱۰} با استفاده از نحو یکی از دو زبان مبدا یا مقصد توانستند بر ضعف این مدل‌ها فائق آیند. هرچند که این روش‌های ابداعی توانستند در بهبود و ارتقاء عملکرد ترجمه‌ی ماشینی مثمرثمر باشند، اما ازآنجا که آنها عموماً دو زبان مبدا و مقصد را توام با یکدیگر تجزیه و تحلیل می‌کنند در مقایسه با نظام‌های ترجمه‌ی ماشینی عبارت‌مبنا، از بار پردازشی و محاسباتی بسیار بالایی برخوردارند. رویکرد دیگری که تلاش کرده است بر این ضعف غلبه کند،

1 - Phrase based
2 - Phrase table

رویکردی است که در آن ترتیب لغوی جملات زبان مبدا باتوجه به قواعد اعمالی بر تجزیه و تحلیل زبان مبدا (هم در هنگام آماده‌سازی و هم در هنگام به‌کارگیری) صورت می‌پذیرد (کلینز و دیگران، {۳} و گنزِل، {۴}).

گنزل {۴} رویکردی به یادگیری خودکار قواعد مرتب‌سازی مجدد ارائه می‌دهد تا بتوان آنرا به عنوان مرحله‌ای پیش‌پردازشی در ترجمه‌ی ماشینی عبارت‌مبنا به‌کار بست. در همین راستا، وی روشی عمومی را مطرح می‌سازد که به تحلیل‌گری نیاز دارد که تنها به زبان مبدا دسترسی دارد و قادر است تا مشکلات موجود بر سر راه مرتب‌سازی مجدد در یک سامانه‌ی عبارت‌مبنا را کاهش دهد. در این روش می‌توان بدون کمک گرفتن از زبانشناس آشنا به زبان مورد بررسی و یا حتی گویشوران آن زبان، قواعد مورد نظر برای مرتب‌سازی مجدد را به‌دست آورد. الگوریتم مطروحه کاملاً قدرتمند بوده و در محیط پر آشوب داده‌های اینترنتی به‌خوبی عمل می‌کند.

ویزوواریان و دیگران {۹} نشان می‌دهند که مرتب‌سازی مجدد نحومبنا روشی موثر در بررسی تفاوت‌های مرتبط با ترتیب لغات در زبان مبدا و مقصد در سامانه‌ی ترجمه‌ی ماشینی آماری به حساب می‌آید. ایشان روشی ساده و خودکار برای فراگیری قواعدی که ترتیب خطی جملات زبان مبدا را به نزدیکترین ترتیب خطی در زبان مقصد تبدیل کند، ارائه می‌دهند. در این روش که تنها از تحلیل درختی زبان مبدا و جهت‌گیری خودکار بهره می‌برد؛ قواعد مطروحه در یک مرحله‌ی پیش‌پردازشی به زبان مبدا اعمال می‌گردند. همین مسئله پیشرفتی عظیم را در ترجمه‌ی ماشینی زبان‌های مختلف به یکدیگر سبب می‌شود.

ناواراتیل و دیگران {۷} به تشریح دو روش مرسوم در ترجمه ماشینی از انگلیسی به آلمانی برای تغییر در ترتیب خطی زبان مقصد پرداخته‌اند. این دو روش از داده‌های دوزبانه و هم‌ترازی خودکار لغات به منظور تغییر در ترتیب خطی زبان مبدا به نحوی که بیشترین شباهت را با زبان مقصد داشته باشد، بهره می‌برند. درحالی‌که، اولین روش صورت بسط یافته‌ی الگوریتم تحلیل‌محوری[1] است که عوامل بافتی را بر تجزیه و تحلیل منطبق می‌سازد؛ روش دوم از یک مدل مشخصه‌بنیاد خطی[2] و حل مسئله‌ی TSP برای تغییر در ترتیب خطی، بهره می‌گیرد. نتایج این بررسی نشان می‌دهد که هر دو روش موجب می‌شوند ترجمه از آلمانی به

1 - Parse-based algorithm
2 - Linear feature-based model

انگلیسی و انگلیسی به آلمانی با کیفیت بهتری صورت پذیرد. بااین‌حال، این روش‌ها در ترجمه از آلمانی به انگلیسی نتایج بهتری را به نسبت ترجمه از انگلیسی به آلمانی نشان می‌دهند.

بااین همه، اکثر پژوهش‌های صورت گرفته پیرامون ترجمه‌ی ماشینی اطلاعات زبانشناختی را نادیده گرفته و صرفاً به اطلاعاتی آماری و ریاضی‌گونه بسنده کرده‌اند. همین مسئله وجه افتراق رویکرد مطروحه در این مقاله با رویکردهای دیگر است. ازآنجاکه، مهمترین تفاوت ظاهری میان زبان انگلیسی و فارسی این است که انگلیسی یک زبان هسته‌ابتدا و فارسی یک زبان هسته‌انتهاست و به‌عبارتی، در انگلیسی در حالت بی‌نشان شاهد ترتیب خطی «فاعل فعل مفعول» هستیم اما در فارسی این ترتیب خطی به صورت «فاعل مفعول فعل» می‌باشد؛ در این مقاله به منظور بهبود عملکرد سامانه‌ی ترجمه‌ی ماشینی، یک سامانه‌ی تغییر ترتیب خطی که سامانه‌ی مرتب‌سازی مجدد[1] نامیده می‌شود، ارائه می‌گردد. اطلاعات مورد نیاز برای این سامانه از تحلیل‌گر استنفورد[2] به‌دست آمده است و برای تبدیل جملات زبان انگلیسی به فارسی و به‌عبارتی، تبدیل یک زبان هسته‌ابتدا به هسته‌انتها بر اساس ویژگی‌های مطروحه از سوی هایدر {۵} و تحلیل صورت گرفته در تحلیل‌گر استنفورد، قواعد و اصولی ا استخراج و به‌کار بسته می‌شوند. سپس این قواعد در قالب یک پردازشگر، به زبان انگلیسی اعمال می‌شوند تا ترتیب خطی نزدیک به زبان مقصد که فارسی است، به دست آید. این سامانه برای شناسایی ویژگی‌های زبان‌های دارای ترتیب‌های خطی متفاوت، آماده‌سازی شده و تعلیم دیده است. بر همین اساس، ترتیب خطی جملات زبان مبدا قبل از ترجمه بر اساس آن اطلاعات زبانی زبان مقصد اصلاح می‌شوند که در سامانه‌ی مرتب‌سازی مجدد گنجانده شده است. از این‌رو، انتظار می‌رود که ترجمه حاصله بیشترین نزدیکی را با زبان مقصد داشته باشد.

بر همین اساس در این مقاله این مسئله مدنظر قرار دارد که نشان داده شود از چه اطلاعاتی می‌توان برای شناسایی ترتیب خطی یک زبان (در اینجا فارسی و انگلیسی) بهره برد و اینکه در نظر گرفتن مرتب‌سازی مجدد در ترجمه‌ی ماشینی چه مزایایی بر در نظر نگرفتن آن دارد. در ادامه ابتدا ترتیب خطی و نحوه‌ی شناسایی و تشخیص آن در زبان‌های انگلیسی و فارسی ارائه می‌گردد و آنگاه با ارائه‌ی فرایند مرتب‌سازی مجدد نشان داده می‌شود که چگونه

1- Reordering system
2 - Stanford parser

می‌توان از راهکارهای شناسایی ترتیب خطی در راستای ترجمه‌ی ماشینی بهره جست.

۲. ترتیب خطی و ترجمه ماشینی

ترتیب‌خطی

به باور هایدر {۵ : ۴} بخش مهمی از تفاوت میان عبارات و بندهای هسته‌ابتدا و هسته-انتها حاصل همراهی «نیازهایِ شناسایی وابسته/ متمم توسط هسته» با «محدودیت ساختی پایه» است و به عبارتی، می‌توان گفت که خطی‌سازیِ۱ هسته و متمم تابع پارامتر جهت‌مندی و شناسایی متمم/وابسته توسط هسته می‌باشد. بر این اساس، جایگاه‌های ادغامی در فرافکن هسته‌ی یک عبارت، می‌بایست با توجه به اصل شناسایی جهت‌مند۲، به‌طور مناسب بوسیله‌ی هسته و برطبق جهت‌مندی بنیادین۳، شناسایی شوند. پیوند محدودیت ساختی پایه و اصل شناسایی جهت‌مند گستره‌ی وسیعی از ویژگی‌های ساختارهای هسته‌ابتدا و هسته‌انتها را به‌دست می‌دهد. هایدر {همان: ۵} نتایج نظریه خود را به صورت زیر خلاصه می‌کند:

"زبان‌های هسته‌انتها (OV) نتیجه‌ی مستقیم پیوند محدودیت ساختی پایه با اصل شناسایی جهت‌مند هستند اما زبان‌های هسته‌ابتدا (VO) نتیجه‌ی مستقیم چنین پیوندی نبوده و به علت ساختارهای پوسته‌ای۴، پیچیده‌تر از زبان‌های هسته‌انتها می‌باشند؛ بااین‌حال، ازآنجاکه هسته را زودتر ارائه می‌کنند، از لحاظ پردازشی دارای مزیت بوده و می‌توانند برای پردازش‌های صعودی۵ مفیدتر باشند. در مجموع می‌توان گفت که هیچ‌یک از دو زبان فوق ضرورتاً برای شرایط استفاده و پردازش ساختارهای دستوری در هنگام تجزیه و تحلیل۶، بهینه نیستند؛ چراکه زبان‌های هسته‌انتها با وجود ساختاری ساده‌تر، هسته را در انتها عرضه می‌کنند و زبان‌های هسته‌ابتدا علی‌رغم عرضه‌ی زودهنگام هسته، دارای ساختاری به مراتب پیچیده‌تر هستند."

هایدر {۵ : ۵} معتقد است VO و OV مکمل هم نیستند و علاوه بر این دو گونه، گزینه‌ی

1- Linearization restriction
2- Principle of directional identification
۳- جهت‌مندی بنیادین عامل پارامتری بنیادینی است که به ترتیب ساختارهای هسته‌پایانی و هسته‌آغازین را تولید می‌کند.
4- Shell structures
5- Bottom-up
6- Parsing

سومی نیز وجود دارد که جهت‌مندی بنیادین تخصیص‌نیافته[1] نامیده می‌شود. از نظر تاریخی این گونه‌ی سوم، در اجداد زبان‌های هندواروپایی (مانند آلمانی) مشاهده می‌شود. به‌عبارتی، در این زبان‌ها شاهد تغییر از یک جهت‌مندی تخصیص‌نیافته به یک جهت‌مندی تخصیص‌یافته هستیم. این تغییر دو گونه‌ی محتمل را دربر داشته است که عبارتند از OV و VO. به اعتقاد هایدر {همان : ۵۹} اگر زبانی جزو گونه‌ی سوم باشد دارای ویژگی‌های زیر است:

الف. گروه حرف‌تعریفِ پیش‌فعلی در ترتیب بنیادین بیشتر درجاست تا اینکه اشتقاقی باشد: اگر زبانی هسته‌ابتدا باشد گروه حرف تعریفِ پیش‌فعلی در آن اشتقاقی است و نه درجا:

۱) **فاعل** (e_j) DP V DP_j

ب. باوجود حرکت پرسشواژه، پرسشواژه‌ی فاعلی می‌تواند درجا بماند: در زبان‌های نوع سوم وOV، فاعل می‌تواند در جایگاهی که ادغام شده است باقی بماند. بنابراین، پرسشواژه‌ی فاعلی به مانند دیگر پرسشواژه‌ها می‌تواند درجا نیز باشد.

پ. قیدها دارای اثرحاشیه‌ای[2] نیستند: اثرحاشیه‌ای ویژگی زبان‌های هسته‌ابتدا است. این ویژگی بازتابی است از محدودیتی علیه عناصر پسافعلی در یک گروه که به‌عنوان سازه‌های قیدی پیش‌فعلی عمل می‌کنند. براین‌اساس، عناصری می‌توانند پیش از گروه‌قیدی پیش‌فعلی ظاهر شوند اما پس از آن نمی‌توانند. به‌عنوان مثال در انگلیسی که زبانی هسته‌ابتدا است این محدودیت بسیار نمایان است (۲):

(۲)[هایدر، همان: ۴ (**4**)]

a. He has [(much more) carefully (**قید**) (*than anyone else)] analyzed it.

b. He has [(much less) often (**قید**) (*than I (thought))] rehearsed it.

اما در زبان‌های نوع سوم و**OV** چنین چیزی رخ نمی‌دهد. همچنان که در نمونه‌های فارسی شاهد آن هستیم:

(۳)

الف) او [(بسیار) با دقت تر [قید] (از هرکس دیگری)] آنرا بررسی کرد.

1- Underspecified canonical directionality

2- Edge effect

ب) او [اغلب (خیلی کمتر) [قید] (از چیزی که من فکر می‌کنم)] ورزش می‌کند.

ت. تنوع ترتیب فعل‌اصلی و فعل‌غیراصلی: ترتیب فعل‌اصلی و فعل‌غیراصلی نشانگر دیگری است که تمایزی میان زبان‌های هسته‌ابتدا و نوع سوم ایجاد می‌کند. جایگاه اصلی فعل‌غیراصلی در هر زبانِ هسته‌ابتدا بدون استثنا پیش از فعل اصلی است (۴). بنابراین، هرگاه در بندی فعل-غیراصلی پس از فعل اصلی ظاهر شود، این بند نمی‌تواند هسته‌ابتدا باشد (۵).

(۴)

a. Kaveh has (فعل کمکی) eaten (فعل اصلی) the apple.

b. Arash had (فعل کمکی) bought (فعل اصلی) a shirt.

(۵)

الف) کاوه سیب را خورده (فعل اصلی) است (فعل کمکی).

ب) آرش لباس خریده (فعل اصلی) بود (فعل کمکی).

هایدر {همان: ۱۰و۶۲} همچنین به بیان تفاوت میان زبان‌های هسته‌ابتدا و هسته‌انتها می‌پردازد که در ادامه به‌طور خلاصه بیان می‌گردد:

الف. زبان‌های هسته‌ابتدا متراکم هستند اما زبان‌های هسته‌انتها فاقد این ویژگی هستند: به‌عنوان‌مثال جملات (۶) همگی در انگلیسی بدساختند چرا که در (۶**a, b**) عنصری که میان فعل و مفعول و یا دو مفعول واقع شده ویژگی مجاورت را تخریب کرده است و در (۶**c**) شاهد فقدان تنوع ترتیب لغات (قلب‌نحوی) هستیم اما معادل آن در فارسی (۷ الف ـ پ) تماماً خوش‌ساخت است:

(۶) [هایدر، همان: ۱۰۷(6)]

a.*[hug *gently* Mary]
b.*[tell Mary *often* jokes]
c.*[buy *the drink* $_{(i)}$ a friend (e $_{(i)}$)]-[buy a friend the drink]

(۷)

الف. [مریم را به *آرامی* بغل‌کن]

ب. [او برای مریم *اغلب* داستان تعریف می‌کند]

پ. [برای دوستش یک نوشیدنی خرید] / [یک نوشیدنی برای دوستش خرید]

ب. جزءفعلی در زبان‌های هسته‌ابتدا تنها بعد از فعل می‌آید اما در زبان‌های هسته‌انتهای دارای این جزءفعلی، همیشه قبل از فعل ظاهر می‌شود: مثلا در انگلیسی **give back** است و در هلندی **back give**.

پ. جایگاه نقشی فاعل در زبان‌های هسته‌ابتدا از لحاظ واژی اجباری است و به‌عبارتی، می‌بایست همیشه پر باشد. از این‌رو، پوچواژه‌ی فاعلی در زبان‌های هسته‌ابتدا اجباری است ولی چنین سازه‌ای در زبان‌های هسته‌انتها مشاهده نمی‌شود. همچنین فاعل در زبان‌های هسته‌انتها می‌تواند به‌راحتی تحت فرایند خروج قرار بگیرد اما در زبان‌های هسته‌ابتدا چنین چیزی ممکن نیست: به‌عنوان‌مثال، جمله‌ی (۸**a**) در صورت نبود فاعل و جمله (۸**b**) درصورت نبود «**it** » در انگلیسی بدساخت هستند و معادل آنها در فارسی (۹) کاملاً خوش‌ساخت می‌نماید چه با فاعل و چه بدون حضور فاعل:

(۸)

a. He/*pro went to the park
b. it/*pro is sunny

(۹)

الف) او/**pro** به پارک رفت

ب) هوا چطوره؟ ← هوا/**pro** آفتابیه

ت. زبانی که قلب‌نحوی را مجاز شمرد، زبانی هسته‌انتهاست و زبانی که چنین چیزی را مجاز نشمرد، هسته‌ابتدا به‌حساب می‌آید. این مسئله نتیجه‌ی مستقیم اصل شناسایی جهت‌مند و به‌طور خاص نتیجه‌ی نیاز متقابل آمریت/تسلط سازه‌ای کمینه است که مسئول ویژگی‌های تراکم معکوس گروه‌های هسته‌ابتدا و هسته‌انتها می‌باشد. به اعتقاد هایدر در مجموعه‌ای هسته‌ابتدا، قلب‌نحویِ درون‌فعلی[1]، مداخله‌گر غیرمُجازی را تولید می‌کند که آمریت/تسلط سازه‌ای کمینه میان فعل و **XP** را نقض می‌کند (۱۰**a**). همچنین قلب‌نحوی به حاشیه‌ی چپ (۱۰**b**) نیز مقصدی را هدف قرار می‌دهد که هسته‌ی فعلی نمی‌تواند آنرا شناسایی کند؛ چراکه، این جایگاه در دامنه‌ی جهت‌مندی بنیادین قرار ندارد. اما در (۱۰**c, d**) هر گروهی که تحت اشراف یک گره در فرافکن گروه فعلی قرار دارد به صورت جهت‌مند توسط گرهِ خواهرش از

1- VP-internal

سمت راست شناسایی می‌شود. بنابراین، کمینگی تحت آمریت متقابل، حفظ می‌شود. علاوه‌براین، حاشیه‌ی چپ گروه فعلی در دامنه‌ی جهت‌مندی بنیادین هسته قرار دارد. این موارد ملزومات یک دستور برای قلب‌نحوی است که هسته‌انتها آنرا مجاز دانسته و هسته‌ابتدا آنرا غیرمجاز می‌شمرد:

(۱۰) [هایدر، ۵ : ۵۸ (**36**)]

a.* $[_{VP}\ \mathbf{V}^{\circ}_{j}\ [YP_{i}\ [XP\ [e_{j}\ e_{i}]]]]$ (VO: YP scrambled VP-internally)
b.* $[_{VP}\ YP_{i}\ [_{VP}\ \mathbf{V}^{\circ}_{j}\ [XP\ [e_{j}\ e_{i}]]]]$ (VO: YP scrambled to the edge of VP)
c. $[_{VP}\ XP\ [ZP_{i}\ [YP\ [e_{i}\ \mathbf{V}^{\circ}]]]]$ (OV: ZP scrambled VP-internally)
d. $[_{VP}\ ZP_{i}\ [XP\ [YP\ [e_{i}\ \mathbf{V}^{\circ}]]]]$ (OV: ZP scrambled to the edge of VP)

از لحاظ شناختی، دستور، یک نظامِ مدیریتِ مکانی به‌حساب می‌آید که مغز برای پردازش زبانی چه در ادراک و چه در تولید به‌کار می‌گیرد. این نظام، الگوریتمی است که ما در هنگام نگاشت از مولفه‌ی یک‌بعدی (خطی) آوایی به مولفه‌ی حداقل دوبعدی (سلسله‌مراتبی) معنا و بلعکس، به‌کار می‌بریم. ساختارهای آوایی حول محور زمان شکل گرفته و نتیجتاً دارای ترتیب‌خطی هستند اما بازنمودهای مفهومی بی‌زمان بوده و ساختارهای پیچیده‌ی دارای ساختار سلسله‌مراتبی را شکل می‌دهند و نحو پلی است میان این‌دو. حال ازآنجاکه به باور بیرویش {۲} نظام حسی‌حرکتی یک نظام پیچیده است که ضرورتاً به الگوهای خطیِ سیگنال‌های زبانی مرتبط بوده، تولید و ادراک را کنترل کرده و می‌تواند تولید و درک سیگنال‌های آوایی را نظم داده و قاعده‌مند سازد؛ طبق نظر پیاتلی‌پالمارینی و دیگران {۸} ترتیب خطی در این نظام مهم و الزامی است و توسط این نظام صورت می‌پذیرد. زبان فارسی باتوجه به ویژگی‌های مطروحه‌ی فوق از سوی هایدر، توسط نظام حسی‌حرکتی به‌عنوان یک زبان هسته‌انتها صورت‌بندی می‌گردد و زبان انگلیسی یک زبان هسته‌ابتدا محسوب می‌گردد.

علاوه بر موارد فوق، می‌توان باتوجه به نمونه‌های (۱۱) و (۱۲) نیز دریافت که موضوع درونی و به‌عبارتی، مفعول در فارسی برخلاف انگلیسی در حالت بی‌نشان پیش از فعل قرار می‌گیرد.

(۱۱)

الف) سهراب سارا را دید.

ب) سهراب دیدسارا را. (نشان‌دار)

پ) آنها کتاب‌ها را سوزاندند.

ت) آنها سوزاندند کتاب‌ها را. (نشان‌دار)

ث) علی حسن را به مدرسه رساند.

ج)علی به مدرسه رساند حسن را. (نشان‌دار)

چ) آرش کتابی خرید.

ح)آرش خرید کتابی. (نشان‌دار)

خ)کاوه لوازمش را فروخت.

د)کاوه فروخت لوازمش را. (نشان‌دار)

(۱۲)

a) He buys a book but	a') * He a book buys
b) John gave Mary a book but	b') *John Mary gave a book
c) John gave a book to Mary but	c') *John a book gave Mary

جملات فوق نشانگر این مطلب هستند که جایگاه موضوع درونی (مفعول) نسبت به محمول (فعل) در فارسی باتوجه به نشان‌دار بودن یا نشان‌دار نبودن تعبیر می‌شود اما در انگلیسی جایگاه موضوع درونی نسبت به محمول با توجه به خوش‌ساختی یا بدساختی تعبیر می‌گردد. به‌عبارتی، قرار گرفتن موضوع درونی یک فعل در فارسی بعد از فعل باعث نشان‌دار شدن ساخت می‌گردد و نه بدساختی آن (نمونه‌های۱۱ ب، ت، ج، ح و د)؛ درحالی‌که، در انگلیسی قرار گرفتن موضوع درونی پیش از فعل به بدساختی جمله منجر می‌گردد (نمونه‌های ۱۲a', b', c'). همچنین در مورد جایگاه دو مفعول مستقیم و حرف‌اضافه‌ای نسبت به فعل، زبانشناسانی همچون کریمی {۶} معتقدند که در فارسی دو جایگاه بی‌نشان برای مفعول مستقیم وجود دارد: الف) اگر مفعول صریح/مستقیم «مشخص» بوده و به‌عبارتی، به‌طور اجباری با «را» همراه شده باشد، مقدم بر گروه حرف‌اضافه‌ای یا مفعول غیرصریح/غیرمستقیم قرار می‌گیرد (۱۳الف و الف') و ب) اگر این مفعول غیرمشخص باشد (و بدون «را» بیاید)، جایگاه بی‌نشان آن پس از گروه حرف‌اضافه‌ای و یا مفعول غیرمستقیم/غیرصریح و در مجاورت فعل است (۱۳ب و ب'):

(۱۳)

الف) «فاعل» «مفعول صریح مشخص» «مفعول حرف‌اضافه» «فعل»

الف') من کتاب را به آریا دادم.

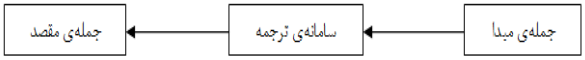
ب) «فاعل» «مفعول حرف‌اضافه» «مفعول صریح غیرمشخص» «فعل»

ب') من به آریا کتاب دادم.

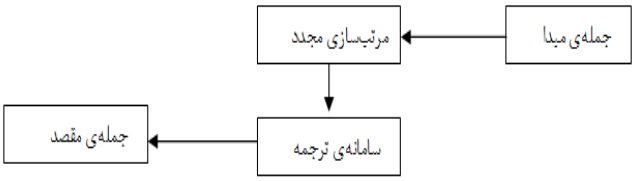
باتوجه به این مسائل و تفاوت ظاهری میان انگلیسی و فارسی، ماشین ترجمه‌ای که بتواند به خوبی ترتیب‌خطی زبان مبدا را به صحیح‌ترین شکل ممکن به ترتیب‌خطی زبان مقصد برگرداند، بسیار پراهمیت و ضروری است و نظامی که بتواند به بهینه‌ترین صورت ممکن ترتیب خطی زبان مبدا را تغییر داده و ترتیبی خطی مشابه با زبان مقصد ارائه دهد، می‌تواند پیشرفت مهمی را در ترجمه از انگلیسی به فارسی رقم بزند.

مرتب‌سازی مجدد

مرتب‌سازی مجدد فرایندی است که هدفش ارائه‌ی ترتیب خطی مشابه با زبان مقصد در ترجمه‌ی ماشینی است. در ترجمه‌ی ماشینی، مرتب‌سازی مجدد لغات که به‌عنوان ابزار پردازشی عمل می‌کند، می‌تواند به آسان‌تر شدن پردازش ماشین ترجمه کمک شایانی کند. پردازش، دسته‌بندی[1]، ریشه‌گیری[2] و موارد مشابه گامی مهم در کاربردهای زبان طبیعی محسوب می‌شوند. مرتب‌سازی مجدد لغات در سطح جمله که مرحله‌ای هزینه‌بر در پردازش زبانی به‌حساب می‌آید در بهبود نتایج حاصل از ترجمه‌ی ماشینی موثر است.

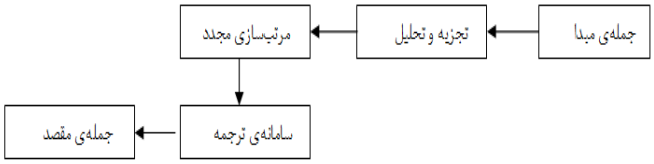


تصویر ۱- سامانه‌ی مرسوم ترجمه



تصویر ۲- سامانه‌ی ترجمه دارای مرتب‌سازی مجدد

1- Tokenization
2- Stemming



تصویر ۳- سامانه‌ی مرتب‌سازی مجدد پیشنهادی

در تصویر شماره‌ی ۱ شاهد سامانه‌ی مرسوم ترجمه هستیم که در آن جمله‌ی مبدا به سامانه‌ی ترجمه داده می‌شود و این سامانه آنرا به زبان مقصد ترجمه می‌کند. برای بهبود ترجمه، سامانه‌ی ترجمه را به مکانیزم مرتب‌سازی مجدد مجهز نمودیم. در تصویر شماره‌ی ۲ سامانه‌ی ترجمه به همراه این مکانیز نشان داده شده است. در اینجا جمله‌ای از زبان مبدا به مکانیز مرتب‌سازی مجدد داده می‌شود و سپس به زبان مقصد ترجمه می‌شود. در تصویر شماره‌ی ۳ نحوه‌ی عملکرد مکانیز مرتب‌سازی مجدد نشان داده شده است. در این تصویر نشان داده می‌شود که جمله ابتدا (به‌طور مثال به‌وسیله‌ی تحلیل‌گر استنفورد) مورد تجزیه و تحلیل قرار می‌گیرد و سپس مجموعه‌ای از گشتارها با هدف تبدیل ترتیب‌خطی جمله‌ی مبدا به آن ترتیب‌خطی که به زبان مقصد نزدیک‌تر باشد، به درخت‌های حاصل از تحلیل نحوی اعمال می‌گردند.

۳. ترجمه میان زبان‌های هسته‌ابتدا و هسته‌انتها

به علت تفاوت ظاهری میان فارسی و انگلیسی از لحاظ ترتیب خطی، بررسی مسئله‌ی مرتب‌سازی مجدد در ترجمه‌ی ماشینی این دو زبان به یکدیگر، بسیار مهم و حیاتی است. برای همین منظور جمله‌ی زیر را در نظر بگیرید:

(۱۴)

He went to shop

عناصر موجود در جمله‌ی (۱۴) به شرح ذیل می‌باشد:

(۱۵)

He: فاعل

Went: فعل

To: حرف‌اضافه

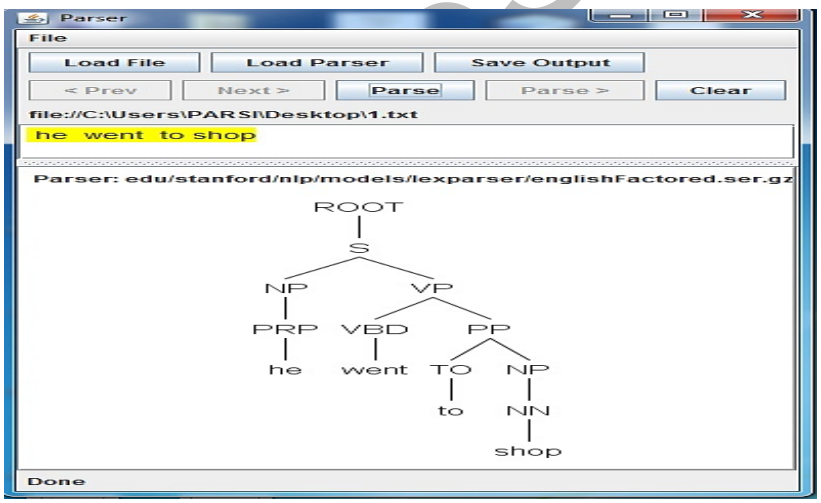
Shop: مفعول

ترتیب‌خطی هسته‌ابتدا در زبان مبدا به ترتیب‌خطی هسته‌انتها در زبان مقصد تبدیل می‌شود. در این نمونه، ترتیب‌خطی جمله‌ی مقصد همانند ترتیب‌خطی جمله‌ی مبدا نیست، بنابراین پس از مرتب‌سازی مجدد صورت خطی جمله‌ی مبدا به صورت زیر خواهد بود:

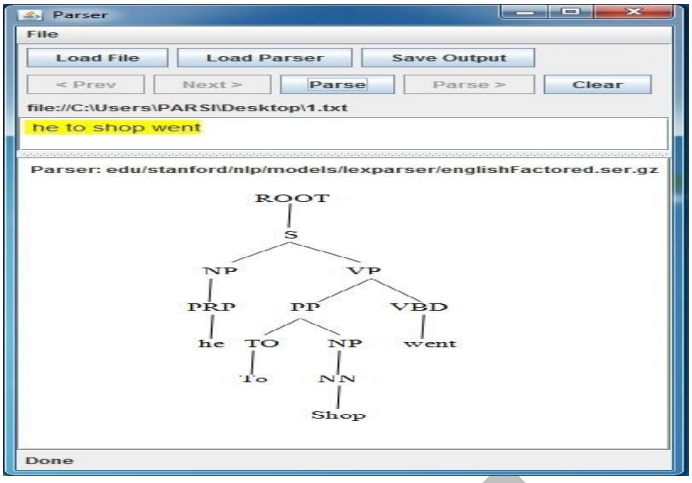
(۱۶)

He went to shop (زبان مبدا) → he to shop went (زبان مقصد)

این گونه از مرتب‌سازی بر اساس ترجمه‌ی انسانی و بر طبق قواعدی خاص صورت پذیرفته است که این قواعد با استفاده از اطلاعات به دست آمده از تحلیل‌گر استنفورد تولید شده‌اند (جدول ۱). ساختار درختی تولید شده برای جمله‌ی زبان مبدا به صورت تصویر۴ و برای جمله‌ی زبان مقصد به صورت تصویر ۵ می‌باشد:



تصویر۴- ساختار درختی زبان مبدا



تصویر۵- ساختار درختی زبان مقصد

جدول۱ تفاوت در چینش اجزای کلام در هنگام ترجمه‌ی این جمله به زبان مقصد و به‌عبارتی، قواعد سازه‌ای دخیل در مرتب‌سازی مجدد را نشان می‌دهد:

جدول ۱: مرتب‌سازی مجدد یک جمله‌ی ساده با توجه به قواعد سازه‌ای

جمله: he went to shop	
S → NP VP VP → VBD PP	زبان مبدا
S → NP VP VP → PP VBD	زبان مقصد

۴. نحوه‌ی عملکرد سامانه‌ی مرتب‌سازی مجدد

اکنون باید دید که نحوه‌ی عملکرد این سامانه‌ی مرتب‌سازی به چه صورت است. همان‌گونه که در ۳-۱ بیان شد، هایدر {۵} زبان‌ها را به سه دسته تقسیم می‌کند: الف) زبان‌های دارای ترتیب‌خطی بنیادین یا پایه، ب) زبان‌هایی با ترتیب خطی هسته‌ابتدا و پ) زبان‌هایی با ترتیب خطی هسته‌انتها. وی برای شناسایی این زبان‌ها ویژگی‌هایی را مطرح می‌سازد که به باور نویسنده‌ی این سطور یک سامانه‌ترجمه‌ی ماشینی هنگامی موفق عمل می‌کند که بتواند ابتدا

به ساکن این ویژگی‌ها را از یکدیگر تمایز داده و بر اساس آن زبان‌ها را طبقه‌بندی نماید. به منظور دستیابی به این هدف، این ویژگی‌ها می‌بایست در ذیل سه مقوله‌ی «ترتیب بنیادین»، «هسته‌ابتدا» و «هسته‌انتها»، برای این سامانه تعریف شوند. سپس با استفاده از هوش‌مصنوعی و پیکره‌های موجود (مخصوصاً پیکره‌های به‌روزشونده) مشخص می‌شود که زبان‌های مورد بررسی ذیل کدامیک از این گروه‌ها قرار می‌گیرد. به‌عنوان مثال، اگر باتوجه به پیکره مشخص شود که ویژگی‌های (۱۷) در زبان مورد بررسی بیشتر از ویژگی‌های (۱۸) خودنمایی می‌کنند، آن زبان در طبقه‌ی هسته‌ابتدا قرار می‌گیرد و اگر ویژگی‌های (۱۸) بر ویژگی‌های (۱۷) اولویت یافت، آن زبان در طبقه‌ی هسته‌انتها واقع می‌شود.

(۱۷) (زبان‌های هسته‌ابتدا)

الف. زبان‌های هسته‌ابتدا متراکم هستند. به‌عنوان‌مثال در این زبان هیچ عنصری (قید) نمی‌تواند میان فعل و مفعول و یا دو مفعول قرار بگیرد و قلب نحوی نیز نمی‌تواند این تراکم را از میان بردارد:

a.*[hug *gently* Mary]
b.*[tell Mary *often* jokes]
c.*[buy *the drink* $_{(i)}$ a friend (e $_{(i)}$)]-[buy a friend the drink]

ب. جزءفعلی در زبان‌های هسته‌ابتدا تنها بعد از فعل می‌آید، مثلا در انگلیسی همیشه شاهد ترکیب give back خوش‌ساخت و ترکیب back give بدساخت است.

پ. جایگاه نقشی فاعل در زبان‌های هسته‌ابتدا از لحاظ واژی اجباری است و به‌عبارتی، می‌بایست همیشه پر باشد. از این‌رو، پوچواژه‌ی فاعلی در این زبان‌ها اجباری است، همچنین فاعل در این زبان‌ها نمی‌تواند به‌راحتی تحت فرایند خروج قرار بگیرد. به‌عنوان‌مثال، جمله‌ی (a) در صورت نبود فاعل و جمله (b) درصورت نبود «it » در انگلیسی بدساخت هستند:

a. He/*pro went to the park
b. it/*pro is sunny

ت. زبانی که قلب‌نحوی را مجاز نشمرد، هسته‌ابتدا به‌حساب می‌آید. به‌عنوان مثال، در انگلیسی جمله (a) خوش‌ساخت و جمله‌ی (b) بدساخت می‌باشد:

a. he gave him a book
b.* him he gave a book (قلب‌نحوی)

ث. در زبان هسته‌ابتدا گروه حرف‌تعریف پیش‌فعلی اشتقاقی است و نه درجا.

ج. پرسشواژه‌ی فاعلی همچون دیگر پرسشواژه‌ها نمی‌تواند درجا باشد و می‌بایست ضرورتاً به جایگاه مشخصگر متمم‌نما حرکت کند.

چ. قیدها دارای اثرحاشیه‌ای[1] هستند؛ براین‌اساس، عناصری می‌توانند پیش از گروه‌قیدی به صورت پیش‌فعلی ظاهر شوند اما پس از آن نمی‌توانند. انگلیسی این محدودیت را نشان می‌دهد:

a. He has [(much more) carefully (*than anyone else)] analyzed it.
b. He has [(much less) often (*than I (thought))] rehearsed it.

ح. جایگاه اصلی فعل‌غیراصلی در هر زبانِ هسته‌ابتدا بدون استثنا پیش از فعل اصلی است. بنابراین، هرگاه در بندی فعل‌غیراصلی پس از فعل اصلی ظاهر شود، این بند نمی‌تواند هسته‌ابتدا باشد. به عنوان نمونه، در انگلیسی جمله‌ی (a) خوش‌ساخت و جمله‌ی (b) بدساخت است:

a. he is playing/ he has studied/ she will open the bag
*b. he playing is / he studied has / she open will the bag

(۱۸) (زبان‌های هسته‌انتها)

الف. زبان‌های هسته‌انتها متراکم نیستند. به‌عنوان‌مثال در این زبان عناصری همچون قید می‌تواند میان فعل و مفعول و یا دو مفعول قرار بگیرد و قلب نحوی نیز می‌تواند این تراکم را از میان بردارد:

الف. [مریم را به *آرامی* بغل‌کرد]

ب. [او برای مریم *اغلب* داستان تعریف می‌کند]

پ. [برای دوستش یک نوشیدنی خرید] / [یک نوشیدنی برای دوستش خرید]

ب. اگر زبانی دارای جزءفعلی باشد و این عنصر همیشه قبل از فعل ظاهر شود، آن زبان هسته‌انتهاست. مثلاً در هلندی همیشه ترکیبback give دستوری است و نه give back.

پ. جایگاه نقشی فاعل در زبان‌های هسته‌انتها از لحاظ واژی اجباری نیست و به‌عبارتی، لازم نیست همیشه پر باشد. از این‌رو، پوچواژه‌ی فاعلی در این زبان‌ها مشاهده نمی‌گردد، همچنین فاعل در این زبان‌ها می‌تواند به‌راحتی تحت فرایند خروج قرار گیرد. به‌عنوان‌مثال، جمله‌ی (الف) و (ب) چه در صورت وجو فاعل و چه در صورت عدم وجود فاعل، کاملاً

1- Edge effect

خوش‌ساخت هستند:

الف) او/pro به پارک رفت

ب) هوا چطوره؟ ← هوا/pro آفتابیه

ت. زبانی که قلب‌نحوی را مجاز شمرد، هسته‌انتها به‌حساب می‌آید. به‌عنوان مثال، در فارسی هر دو جمله‌ی (الف) و (ب) خوش‌ساخت هستند:

الف) این کتاب را به او دادم.

ب) به او این کتاب را دادم. (قلب‌نحوی)

ث. در زبان هسته‌انتها گروه حرف‌تعریف پیش‌فعلی اشتقاقی است و نه درجا.

ج. پرسشواژه‌ی فاعلی همچون دیگر پرسشواژه‌ها می‌تواند درجا باشد و نیاز نیست ضرورتاً به جایگاه مشخصگر متمم‌نما حرکت کند.

چ. قیدها دارای اثرحاشیه‌ای نیستند؛ براین‌اساس، عناصری که می‌توانند پیش از گروه-قیدی به صورت پیش‌فعلی ظاهر شوند اما پس از آن نیز می‌توانند. به‌عنوان مثال در فارسی شاهد این محدودیت نیستیم:

الف) او [(بسیار) با دقت تر (از هرکس دیگری)] آنرا بررسی کرد.

ب) او [اغلب (خیلی کمتر) (از چیزی که من فکر می‌کنم)] ورزش می‌کند.

ح. جایگاه اصلی فعل‌غیراصلی در هر زبانِ هسته‌انتها هم پیش از فعل اصلی است و هم پس از آن. به عنوان نمونه، فعل غیراصلی در فارسی در جمله‌ی (الف) به صورت پیش فعلی و در جمله‌ی (ب) به صورت پس‌فعلی ظاهر شده است:

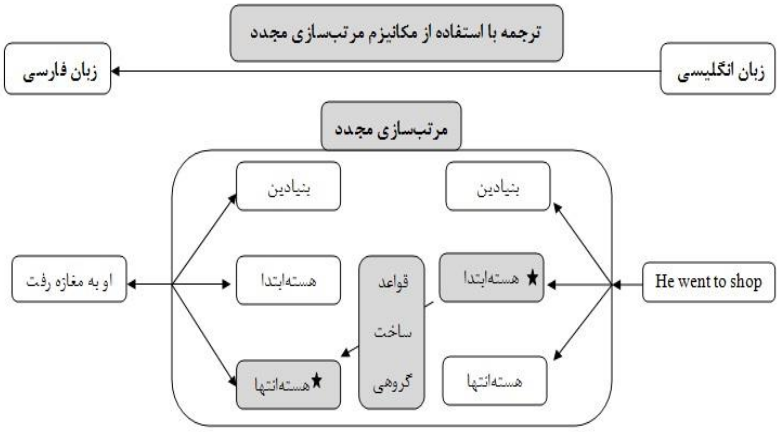
الف) او خواهد رفت/ این را می‌توان گفت/ این‌گونه می‌شود رفت

ب) او رفته است/ من گفته بودم/ او کشته شد

پس از این مرحله و پس از برچسب‌گذاری اجزای سخن، این سامانه باتوجه به قواعد ساخت‌گروهی که در جدول ۱ نشان داده شد، می‌تواند از زبان مبدا به زبان مقصد ترجمه نماید. با چنین سامانه‌ای می‌توان تضمین کرد که ترتیب خطی ترجمه‌ی انجام شده بیشترین نزدیکی را با ترتیب خطی زبان مقصد داشته باشد.

بر اساس آنچه که گفته شد اگر جمله‌ای از زبان X در اختیار این سامانه‌ی ترجمه قرار گیرد و به شرط آنکه این سامانه پیشتر باتوجه به پیکره‌های زبانی، طبقه‌ی زبان X را مشخص

کرده باشد و آنرا در ذیل یکی از سه گروه ترتیب بنیادین، هسته‌ابتدا و یا هسته‌انتها قرار داده باشد، آنگاه باتوجه به ترتیب خطی زبان مقصد، ترتیب خطی جمله‌ی مورد نظر را تغییر داده و آنرا به آن ترتیب خطی که به زبان مقصد نزدیک باشد، برگرداند. مکانیز چنین فرایندی را در تصویر ۶ مشاهده می‌کنید.



تصویر۶- ترجمه همراه با مرتب‌سازی مجدد

۵. نتیجه‌گیری

تغییر ترتیب‌خطی، فرایند موثر و مهمی است که به‌طور معناداری عملکرد ترجمه‌ی ماشینی را بهبود می‌بخشد. به علت اینکه زبان‌های مختلف، ترتیب‌خطی متفاوتی را به کار می‌گیرند، یکی از نیازهای ترجمه‌ی ماشینی آن است که لغات مقصد را در ترتیبی صحیح قرار دهد. درحالی‌که، سامانه‌های ترجمه‌ی ماشینی عبارت‌مبنا در محدوده‌ی کوچکی از لغات و به عبارتی، در محدوده‌ی جملات کوتاه به درستی ترتیب‌خطی را تغییر می‌دهند، برای تغییر صحیح ترتیب‌خطی در جملات بلندتر، با چالشی بزرگ مواجه هستند. ازاین‌رو، در این مقاله به منظور بهبود عملکرد سامانه‌ی ترجمه‌ی ماشینی، سامانه‌ی تغییر ترتیب خطی که سامانه‌ی مرتب‌سازی مجدد نامیده می‌شود، ارائه گردید. در این سامانه‌ی پیشنهادی با استفاده از اطلاعات به‌دست آمده از تحلیل‌گر استنفورد، برای تبدیل جملات زبان انگلیسی به فارسی و به‌عبارتی، تبدیل یک زبان هسته‌ابتدا به هسته‌انتها، بر اساس ویژگی‌های مطروحه از سوی

هایدر {۵} و تحلیل صورت گرفته در تحلیل‌گر استنفورد، ابتدا قواعد و اصولی استخراج و مورد استفاده قرار گرفتند. سپس این قواعد در قالب یک پردازشگر، به زبان انگلیسی اعمال گردید تا ترتیب خطی نزدیک به زبان مقصد که فارسی است، به دست آید.

حال ازآنجاکه پژوهش‌های متفاوت نشان داده‌اند که پردازش، موثرترین روش در به دست آوردن صحیح‌ترین ترتیب‌خطی به‌حساب می‌آیند که با توالی خطی در زبان مقصد بیشترین انطباق را داشته باشد؛ رویکرد مطروحه، شامل یک مرحله‌ی پیش‌پردازشی است که در آن سامانه‌ی مرتب‌سازی مجدد برای شناسایی ویژگی‌های زبان‌های دارای ترتیب‌های خطی متفاوت، آماده‌سازی شده و تعلیم دیده است. بر همین اساس، ترتیب خطی جملات زبان مبدا پیش از ترجمه بر اساس آن اطلاعات زبانی زبان مقصد اصلاح می‌شوند که در سامانه‌ی مرتب‌سازی مجدد گنجانده شده‌اند. مدل پیشنهادی روش موثری را برای تغییر ترتیب‌خطی زبان مبدا بر طبق ویژگی‌های زبان مقصد ارائه می‌کند. با این بررسی می‌توان نشان داد که استفاده از دانش زبان‌شناسی در پردازش داده‌های مورد بررسی، می‌تواند به پیشرفت‌های قابل توجهی در عملکرد ترجمه منتهی گردد.

منابع

[1] Al-Onaizan, Y. and Papineni, K., 2006. "Distortion models for statistical machine translation" In Proceedings of Association for Computational Linguistics.

[2] Bierwisch, M. , 2007. "Semantic form as interface". *Interfaces and interface conditions*, *132*.

[3]Collins, M., Koehn, P. and Kucerova, I., 2005."Clause restructuring for statistical machine translation" In Proceedings on Association for Computational Linguistics, p. 531-540.

[4] Genzel, D., 2010. "Automatically learning source-side reordering rules for large scale machine translation," In Proceedings of the 23rd International Conference on Computational Linguistics.

[5] Haider, H. , 2013. *Symmetry breaking in syntax* (No. 136). Cambridge University Press.

[6]Karimi, S. , 2005. *A minimalist approach to scrambling: Evidence from Persian* (Vol. 76). Walter de Gruyter.

[7] Navratil, J., Visweswariah, K., & Ramanathan, A., 2012. "A comparison of syntactic reordering methods for english-german machine translation". *Proceedings of COLING 2012*, 2043-2058.

[8] Piattelli-Palmarini, M., & Vitiello, G., 2015. Linguistics and some aspects of its underlying dynamics. *arXiv preprint arXiv:1506.08663*.

[9] Visweswariah, K., Navratil, J., Sorensen, J., Chenthamarakshan, V. and Kambhatla, N. , 2010."Syntax based reordering with automatically derived rules for improved statistical machine translation," In Proceedings of the 23rd International Conference on Computational Linguistics.

[10] Yamada, K. and Knight, K., 2002. "A decoder for syntax-based statistical machine translation" In Proceedings of Association for Computational Linguistics.

تحلیل آماری واژه‌های فارسی مقالات علوم انسانی بر مبنای قانون زیف

نجمه امینی‌خواه*، محمدباقر دستغیب و محمدرضا فلاحتی قدیمی فومنی*****

چکیده

در پی کمبود ابزارهای ابتدایی پردازش زبان طبیعی فارسی و نیاز روزافزون به برنامه‌های ماشینی مبتنی بر زبان طبیعی، با مطالعه و اثبات تابعیت زبان فارسی از قوانین زبانشناسی کمّی، می‌توان بین زبان‌های برنامه‌نویسی و زبان‌های طبیعی پل ارتباطی ایجاد کرد. قانون زیف از جمله قوانینی است که در زبانشناسی کمی، در عین سادگی می‌تواند نقشی مهمی در پردازش زبان طبیعی فارسی ایفا کند. چرا که با استفاده از نتایج و گزارش‌های حاصل از این تحلیل، می‌توان برنامه‌ها و ابزارهای پردازش زبان طبیعی را به گونه‌ای اصولی‌تر ساخت. در پژوهش حاضر پیکره‌ای کوچک مقیاس، ساخته شد و سطح قابل قبولی از پیش‌پردازش با رویکردی زبانشناسانه بر روی آن اجرا و سپس قانون زیف بر روی آن پیاده‌سازی و بردار و نمودارهای زیف آن رسم شد. به منظور اعتبارسنجی از ضریب همبستگی پیرسون ما بین بسامدهای تخمینی و واقعی استفاده شد. همچنین بردارهای رسم شد با بردارهای زیف پیکره‌های دیگر به زبان انگلیسی که از این قانون پیروی می‌کنند مقایسه شد. نتیجه حاصل، تبعیت زبان فارسی از این قانون بود.

واژه‌های کلیدی: قانون زیف، پیکره‌های زبانی، آمار، بسامد، زبان فارسی

۱- مقدمه

در سال ۱۹۳۵ زبانشناس آمریکایی جورج کینگزلی زیف[1] دریافت که در زبان طبیعی،

---

* دانشجوی کارشناسی ارشد در رشته زبانشناسی رایانشی در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، n.aminikhah@yahoo.com

** استادیار گروه پژوهشی طراحی و عملیات سیستم‌ها در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری (نویسنده مسئول)، dastghaib@ricest.ac.ir

*** استادیار گروه پژوهشی زبانشناسی رایانشی در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، mrfalahat@yahoo.com

1 George Kingsley Zipf

میان طول کلمات و میزان بسامد آنها ارتباط معکوسی وجود دارد. یکی از نظریات مطرح در رفتار انسان، اصل کمترین کوشش زیف است که در سال ۱۹۴۹ ارائه شد. این اصل حاکی از آن است که انسان تمایل دارد در حل یک مسأله راهی را برگزیند که کمترین تلاش را نیاز دارد. کانوال[۱]، اسمیت[۲]، کول برستون[۳] و کربی[۴] (۲۰۱۷)[1] اذعان دارند که زیف مشاهدات کلاسیک را در خصوص با رابطه بین طول کلمه و بسامد آن ارائه و بیان کرد که یک کلمه با بسامد بیشتر، طول کمتری دارد و همچنین ادعا کرد، این «قانون اختصار» یک ویژگی ساختاری جهانی زبان است. از آن زمان قانون اختصار در طیف گسترده‌ای از زبان های انسانی اثبات شده است و به سیستم‌های ارتباطی حیوانات و حتی زبان‌های برنامه‌نویسی کامپیوتری نظیر جاوا اسکریپت[۵] و سی‌پلاس‌پلاس[۶] گسترش یافته است. مشاهده شده است که توزیع بسامد در موسیقی، جمعیت شهری، انقراض، زلزله و حتی دی‌ان‌ای[۷] نیز براساس قانون زیف است و این توزیع به عنوان محیط‌های زیفی[۸] شناخته می‌شود. .مشاهدات کمّی و مدل‌های ریاضی اولیه‌ی در حوزه‌ی واژگان توسط استوپ (۱۹۱۶)، یول (۱۹۲۴) و کاندن (۱۹۲۸) انجام گرفت. اما جرج کینگلی زیف (۱۹۵۰-۱۹۰۲) بود که رابطهٔ بین کلمات و میزان بسامد را به‌طور قانونمند بررسی کرد. او نخستین کسی بود که یک مدل نظری برای توضیح این روابط یافت و در این رابطه یک فرمول ریاضی ارائه کرد که به « قانون زیف» شهرت یافت.

در این پژوهش هدف، بررسی آماری واژگان زبان فارسی براساس قانون زیف می‌باشد و اساساً این نتیجه مدنظر است که آیا این قانون بر روی زبان فارسی نیز حاکم است یا خیر؟ اما این مهم زمانی به درستی نتیجه خواهد داد که داده‌ها پیش‌پردازش شده و نرمال باشند و چالش‌های موجود تا حد امکان در نظرگرفته شوند.

بررسی تاریخچه پژوهش نشان می‌دهد که پژوهشگران بسیاری به تحلیل زبان‌های مختلف براساس اصل قانون زیف پرداخته‌اند. براساس این پژوهش‌ها، قانون زیف بر روی بیشتر زبان‌ها

---

1 Jasmeen Kanwal
2 Kenny Smith
3 Jennifer Culbertson
4 Simon Kirby
5 Javascript
6 C++
7 DNA
8 Zipfian

نظیر انگلیسی، فرانسه، یونانی، رومی، آلمانی و ... قابل اعمال است.[1] این خود نقطه الهام بخشی بود که بتوان با استفاده از قانون زیف در علوم زبانی و رایانه‌ای، پلی ایجاد کرد و شبیه‌سازی زبان طبیعی برای سیستم‌ها و ماشین‌ها را انجام داد.

از آنجا که زبانشناشی رایانشی رشته‌ای نوپا در زبان فارسی است، انجام پردازش‌های سیستمی برای این زبان از این منظر بسیار کم است. در پژوهش پیش‌رو سعی بر این است که تابعیت زبان فارسی از قانون زیف که یکی از قانون‌های پایدار و مبنایی در تحلیل‌های آماری است مطالعه و اثبات شود چرا که براساس نتایج گزارش‌های این تحلیل، می‌توان برنامه‌ها و ابزارهای پردازش زبان طبیعی را اصولی‌تر ساخت. پس می‌توان گفت که این اقدام گام نخستی برای پردازش زبان طبیعی به حساب می‌آید. زیرا متون زبانی برای سیستمی و ماشین‌خوان شدن باید مراحل پرفراز و نشیبی را طی کنند، که مهمترین آنها پیش پردازش است که به دلیل رسم‌الخط متفاوت از انگلیسی، ننوشته شدن مصوتهای کوتاه در فارسی و وجود نیم‌فاصله که مرزبندی بسیار چالش برانگیز در زبان فارسی به وجود می‌آورد از اهمیت بالایی برخوردار است. با استناد بر درودی[1] و دیگران (۲۰۰۴)[2] در ساخت پیکره‌های مدرن فارسی برای یکدست کردن پیکره، به قانون زیف نیاز است. در این پژوهش زبان فارسی را براساس تعداد حروف کلمات بررسی کرده‌اند. با توجه به نتایج و جداول آورده شده به نظر می‌رسد که نرمال‌سازی به خوبی اعمال نشده است و درمواردی همچون «ها» و «های» به عنوان واژه‌های به ترتیب دو حرفی و سه حرفی یاد شده‌است، درحالی که به تنهایی معنای مستقلی را نمی‌رسانند و واژه محسوب نمی‌شوند و در طبقه پسوندهای صرفی قرار دارند. برای محقق شدن این هدف به مراحل پیش‌پردازش نیاز است که ابتدایی ترین آنها نرمال‌سازی است. بدین گونه که فاصله‌ی بین پسوندها و پیشوندهای صرفی یا حتی اشتقاقی به نیم‌فاصله تغییر پیدا کنند. در پژوهش پیش‌رو سعی بر آن شد که نرمالسازی بر روی دیتاهایی که پیکره حاضر را ساخته‌اند انجام شود.

با استفاده از این تحلیل آماری، فهرستی از واژه‌های پرکاربرد که در بیشتر حالات، کم اهمیت‌ترین واژه‌های یک متن را تشکیل می‌دهند به دست می‌آید که این خود در اکثر سطوح پردازش زبان طبیعی حائز اهمیت است. زیرا با حذف واژه‌های‌مانع و توجه به واژگان کلیدی به

1 Darrudi

منظور بازیابی اطلاعات، رسیدن به هدف سریعتر و با دقت بیشتری انجام خواهد پذیرفت.

از جمله پژوهش‌های انجام شده بر روی زبان‌های مختلف می‌توان از زبان عربی که دارای الفبا و رسم‌الخط نزدیک به زبان فارسی است نام برد. در این راستا مقالات عبدالعلی[1]، کوی[2] و سلیمان[3] (۲۰۰۵)[3] گودر[4] و دی رووک[5] (۲۰۰۱)[4] در زبان عربی قابل ذکر است که در وهله‌ی نخست به نرمال‌سازی و تنظیم مجدد پیکره برای آماده‌سازی و استفاده در نرم‌افزارهای آماری پرداخته‌اند و دوم آن را با زبان انگلیسی مقایسه کرده و به بررسی چالش‌های زبان عربی پرداخته‌اند. در فارسی نیز هاشم‌زاده، نخعی و مرادی‌مقدم (۱۳۹۲)[5]، مهدوی نسب (۱۳۹۲)[6]، ترابی (۱۳۸۹)[7]، غروی‌قوچانی(۱۳۸۵)[8]، درودی، حجازی و ارومچیان (۲۰۰۴)[2]، تقی‌یاره، درودی، ارومچیان و انگشتری (۲۰۰۳)[9] و مهری[6] و جماعتی[7] (۲۰۱۷) [10] از این قانون نیز استفاده کرده‌اند و به بررسی دادگان فارسی پرداخته‌اند اما در هرکدام خلاء عدم‌وجود نرمال‌سازی و پیش‌پردازش متن فارسی وجود دارد و چالش‌های زبان فارسی مورد توجه قرار نگرفته‌است.

از جمله زبان‌هایی که قانون زیف در آن بررسی شده است زبان ماندارین چینی است. با استناد بر لین لیو[8]، ژانگ[9]، گنگ[10]، لینگ لایی[11] و وانگ[12] (۲۰۱۷)[11] نویسه‌های چینی واحدهای پایه‌ای برای کلمات چینی هستند و یک کلمه چینی می‌تواند شامل یک، دو یا چند کاراکتر باشد. بسیاری از کاراکترها می‌توانند به عنوان کلمات در زبان چینی عمل کنند. واژه‌هایی که شامل نویسه‌های دوتایی، سه‌تایی و بیشتر می‌شوند، به عنوان بایگرم، ترایگرم و به طور کلی n-gram نامگذاری می‌شوند. در زبان چینی واژه‌ها را با فاصله مانند زبان انگلیسی جدا نمی‌کند، بنابراین یک خواننده باید یک رشته کاراکتر را به کلمات تبدیل کند تا متن‌های

1 Ahmed Abdelali
2 James Cowie
3Hamdy S. Soliman
4Goweder
5 De Roeck
6 Mehri
7 Jamaati
8 Chao-Lin Liu
9 Shuhua Zhang
10 Yuanli Geng
11 Huei-ling Lai
12 Hongsu Wang

چینی را درک کند. ماندارین چینی طی هزاران سال گذشته تکامل یافته است. اسناد نوشته شده در زبان چینی در حال حاضر شامل تعداد زیادی بایگرم و ترایگرم است در حالی که متون کلاسیک چینی تعداد بسیار زیادی یونیگرم را شامل می‌شود. در پژوهش‌های پیشین بر روی زبان چینی اساساً بر روی تطبیق توزیع زیفی بر پیکره‌های چینی تمرکز شده است. نتیجه بدست آمده از نمودارها و منحنی‌های زبان چینی، تطابق آنها با نمودارها و منحنی‌های زیف بود.

اینکه این قانون یک قانون جهانی است و بر روی بیشتر زبان‌ها قابل اعمال است را می‌توان از پژوهش بنتز[1] و فررکانچو[2] (۲۰۱۶)[12] دریافت. آنها قانون اختصاری زیف را در تمام ۱۲۶۳ متن و ۹۸۶ زبان مورد آزمایش قرار داده‌اند. قدرت قانون نیازمند توضیح نظری است و این موضوع اساساً مهم است، زیرا می‌تواند دریچه جدیدی در بحث در مورد جهانی‌های زبان باز کند. زیرا در ادامه می‌توان به وجود خواص جهانی زبان دست یافت. با این حال، ممکن است که جهانی‌های زبان از اصول اساسی انتقال اطلاعات، به جای زبان و تعصبات خاص انسان، بدست آید.

۲- روش پژوهش

برای انجام این پژوهش پیکره‌ای مورد نیاز است که پیش‌پردازش‌های منحصر به زبان فارسی بر روی آن اعمال شده باشد و یا حداقلِ چالش‌های زبان فارسی از قبیل رعایت اصل نیم‌فاصله در پسوندهای تصریفی اضافه شونده، در آخر کلمات و نیز افعال دارای پیشوند و پسوند، در آن مورد توجه قرار گرفته‌باشد. به دلیل عدم وجود و یا در دسترس نبودن چنین پیکره‌ای، ساخت پیکره‌ی مورد نظر با شرایط ذکر شده، از الزامات این پژوهش است. اما همانطور که پیشتر نیز گفته شد، این تصحیحات و برطرف کردن مشکلات متنی نگارشی زبان فارسی، مستلزم وجود نرم‌افزاری جامع است که تمامی گزینه‌های پیش‌پردازشی متون فارسی را دارا باشد. بنابراین ایجاد این نرم‌افزار در اولویت کار قرار گرفت.

برای شروع ی تحلیل به جامعه و نمونه آماری نیاز است. جامعه پژوهش حاضر متون

1 Christian Bentz
2 Ramon Ferrer-i-Cancho

نوشتاری رسمی زبان فارسی است و نمونه برگرفته از آن که از روش تصادفی ساده انتخاب شده است. نمونه انتخابی ۳۵۰ مقاله، از مقالات علوم انسانی، سه کتاب داستان‌بلند (رمان) که یک کتاب از سری مجموعه هری‌پاتر و دو کتاب از مجموعه داستان ارباب حلقه‌ها و همچنین نزدیک به هزار چکیده اخبار است، که واژگان آن مورد تحلیل و بررسی قرار گرفته‌اند.

در این پژوهش ابزارهای مختلفی به شرح زیر استفاده شد. نخستین نرم‌افزار، نرم‌افزار نرمال‌سازی و تجزیه کننده متن[1] است. کار آن پیش‌پردازش متون و تقطیع آنها به توکن[2] و همچنین گرفتن بسامدهای آنها است. با توجه به محاسبه بسامدهای توکن‌های به‌دست‌آمده، خروجی این نرم‌افزار به شکل انواع[3] و بسامد است.

دومین ابزارکار نرم‌افزار صفحه گسترده ی ماکروسافت اکسل [4] است. در راستای هدف پژوهش که بررسی و تحلیل آماری واژگان زبان فارسی بر مبنای قانون زیف است، می‌بایست خروجی حاصل از نرم‌افزار نخست را در اکسل ذخیره کنیم. این عمل به منظور اعمال بهتر فرآیندهای آماری مورد نیاز، صورت می‌گیرد. همچنین یکی از بهترین نرم‌افزارهایی است که فرمت ذخیره شده آن به آسانی توسط نرم‌افزار متلب[5] فراخوانده می‌شود و داده‌های ذخیره شده در آن مورد تحلیل قرار می‌گیرند.

سومین ابزار، نرم‌افزار متلب است. که اجرای تمام عملیات ریاضی و آمار در آن به سادگی انجام می‌شود و به دلیل داشتن اکثر توابع، نیازمند برنامه‌نویسی کمتری است. همچنین کدهای متلب نیز در مباحث علمی از اعتبار بالاتری نسبت به دیگر نرم‌افزارهای برنامه‌نویسی برخوردار است.

در هر متن موارد جالبی برای نمایش وجود دارد، از جمله این موارد می‌توان به طبقه‌بندی و پردازش آن در سطحی پایین اشاره کرد. اگر متن به عنوان یک فهرست از کلمات در نظر گرفته شود، پرسش‌هایی ایجاد می‌شود مانند: شایع ترین کلمات (انواع) در متن چیست؟ تعداد هر توکن چقدر است؟ نقش کلماتی که درصد زیادی از متن را شامل شده‌اند چیست؟ برای پاسخ به این چنین پرسش‌هایی و نشان دادن کلمات تشکیل‌دهنده‌ی هر متن و مشخص کردن

1 Tokenizer
2 Token
3 Types
4 Excel
5 Matlab

نقش‌های آنها نیاز به بررسی این موارد در پیکره‌هایی با مقیاس‌های قابل توجه است. همچنین از نتایج حاصل می‌توان برای پوشش دستورهای زبانی آماربنیاد که در پی جمع‌آوری متون به دست می‌آیند نیز استفاده کرد.

این موضوع که، توزیع بسیار نامتناسب کلمات در متون امکان دارد، با مباحث آماری مورد بررسی قرار می‌گیرند. در بیشتر تحلیل‌ها نشان داده شده است که نزدیک به ۵۰ درصد از متن را کلمات پر بسامد تشکیل می‌دهند که رتبه ۱ تا ۱۰۰ را به خود اختصاص داده‌اند و از سمتی دیگر باز نزدیک به ۵۰ درصد از کلمات دارای بسامد نزدیک به ۱ تا ۲ هستند که این مجموعه از کلمات به ندرت به کار می‌روند. از سویی دیگر می‌توان گفت که بیش از ۹۰ درصد از انواع کلمات ۱۰ بار یا کمتر رخ می‌دهند که میزان دقیق آن برای محدود کردن لیست واژگان، با فرمولهای خاص محاسبه می‌شود.

در پژوهش پیش‌رو از قانون زیف برای تحلیل آماری کلمات متون فارسی در پیکره کوچک مقیاسی که با اندازه‌ای استاندارد برای این منظور، گردآوری شده، استفاده شده‌است. همانطور که در قبل نیز عنوان شد، برای اجرای این قانون نیاز به بسامد کلمات و رتبه‌ی آنها است که بسامد واقعی براساس سیر صعودی به نزولی مرتب شده‌است. رابطه‌ی بین بسامد و رتبه رابطه‌ای عکس است:[13]

$$f \propto \frac{1}{r} \qquad (۱)$$

براساس این قانون باید بسامد نسبت به رتبه تخمین زده شود که این بسامد از رابطه‌ی زیر حاصل می‌شود.

$$f \cdot r = k \qquad (۲)$$

k در این جا یک عدد ثابت است و از رتبه‌ای به بعد اعداد نزدیک به هم را نشان می‌دهد، که با استفاده از آن می‌توان بسامد تخمینی را به‌دست آورد. برای مثال در این پژوهش در قسمت پیکره‌ی کل کلمات، k=190000 بود که با رابطه‌ی $\frac{190000}{r}$ بسامد تخمینی به دست آمده است. نمودار حاصل از این بسامد و بسامد واقعی نموداری است که طبق قانون زیف داده‌ها را به نمایش می‌گذارد. برای این منظور نیز از قانون توان[1] استفاده شد.

1 power law

در قانون توان با استفاده از نمودار لگاریتمی، لگاریتم مبنای ۱۰ بسامد واقعی و رتبه بر روی نمودار کشیده شد وبه وسیله‌ی خطی مستقیم با شیب ۱- نقطه‌ها به هم وصل شد. هر چه که شیب خط به ۱- نزدیکتر باشد، تابعیت داده‌ها از این قانون بیشتر است در نتیجه پیکره از قانون پیروی می‌کند.

برای اعتبارسنجی کار مقایسه‌ای بین نمودار زیف حاصل از پیکره‌ی کل کلمات و نمودار پیکره brown انگلیسی، همچنین نمودار پیکره‌ی کلمات مجموعه سه کتاب داستان با نمودار کتاب داستان آلیس در سرزمین عجایب انجام شد و نتیجه، نزدیک بودن آنها به هم، و تابعیت ۹۶ درصدی پیکره‌های مورد مطالعه در این پژوهش بود.

به منظور سنجیدن همبستگی بسامد تخمینی با بسامد واقعی نیز ازضریب همبستگی پیرسون استفاده شد. در این روش نیز هرچه نتیجه به یک نزدیکتر باشد نشان دهنده همبستگی بالای دو متغیر است. باید گفت که ضریب همبستگی میان بسامد مجموع پیکره‌های مورد مطالعه پژوهش پیش‌رو (۰/۹۶) شد. نتیجه‌گیری به‌دست آمده از این تحلیل‌ها تابعیت زبان فارسی از قانون زیف را نشان می‌دهد.

گفته شد که برای انجام این پژوهش در نخست نیاز به متن خالص مقالات فارسی در حوزه‌ی علوم انسانی بود. برای ایجاد پیکره‌ای کوچک مقیاس، با دریافت حجم قابل ملاحظه‌ای از مقالات از پایگاه مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری و بررسی آنها، تعداد مقالاتی که امکان تبدیل شدن به متن خالص را داشتند انتخاب شد. در مرحله‌ی دوم به نرمال‌سازی متون مقالات در سطح حذف علامات نگارشی، بررسی و اعمال نیم‌فاصله در جایگاه‌های مورد نیاز، حذف اعداد انگلیسی و فارسی و همچنین حذف کاراکترهای غیر زبانی و الفبای انگلیسی پرداخته شد. مرحله سوم گرفتن بسامد از کلمات درون متن مقالات بود. پس از گرفتن بسامد، کلمات به ترتیب از بیشترین به سمت کمترین بسامد در فهرست قرار گرفتند و به عبارتی رتبه‌ی هر کلمه تعیین شد، مرحله‌ی چهارم اعمال قانون زیف بر روی پیکره‌ی ایجاد شده از متون مقالات بود و پس از تحلیل‌های انجام شده نمودار زیف آنها رسم شد. از آنجا که دسترسی به مقالات با محدودیت‌هایی نظیر: عکس بودن متن مقاله و نبود نرم‌افزار تبدیل عکس به متن فارسی، بهم ریختن شکل نوشتاری-الفبایی در مرحله تبدیل اسناد به اسناد متنی و همچنین داشتن حفاظ‌های امنیتی به منظور حفظ محتوای اسناد روبرو بود. همچنین به منظور ایجاد

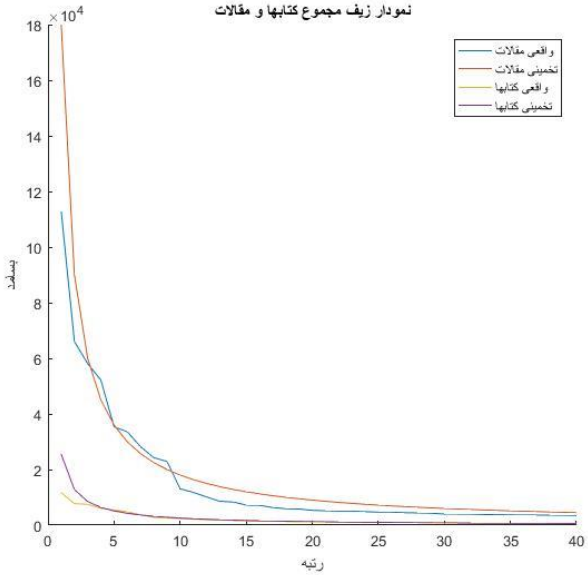
فضای رقابتی، نتیجه کار حاصل از متون استخراج شده از مقالات با متون فارسی در دیگر حوزه‌ها نیز با انجام مراحل ذکر شده بر روی سه کتاب داستان بلند (رمان) از جمله: یکی از کتابهای سری داستان‌های هری‌پاتر و دو کتاب از سری کتابهای ارباب‌حلقه‌ها و همچنین مجموعه‌ای از متن چکیده اخبار نیز مقایسه شد و در آخر نیز تمام پیکره‌ها با هم ادغام شد تا نشان دهد آیا این قانون بر روی متون کلی و ترکیبی از حوزه‌های مختلف نیز پا برجاست یا خیر.

ایجاد فضای رقابتی بین متون را می‌توان به عنوان نقطه‌ی قوت این پژوهش در نظر گرفت، کاری که در ریاضی برای امتحان قضایای به منظور یافتن مثال نقض و رد قضیه یا نظریه انجام داده می‌شود و همچنین بررسی اثبات قانون در حوزه‌های مختلف می‌تواند ابعاد کاربردی این پژوهش را افزایش دهد و از تک بعدی بودن نتیجه دوری کند.

جدول ۱و ۲ و نمودار زیر نتایج حاصل از تحلیل‌های آماری انجام شده در این پژوهش را نشان می‌دهد.

جدول ۱. بسامد واژگان پر تکرار در پیکره‌ی مقالات

رتبه	انواع کلمه	بسامد واقعی	بسامد تخمینی	لگاریتم ۱۰ بسامد واقعی	لگاریتم ۱۰ رتبه
1	و	۱۱۲۸۸۶	۱۸۰۰۰۰	۵.۰۵	۰
2	در	۶۶۰۱۲	۹۰۰۰۰	۴.۹	۰.۳
3	از	۵۸۰۹۲	۶۰۰۰۰	۴.۸	۰.۵
4	به	۵۲۲۰۰	۴۵۰۰۰	۴.۷	۰.۶
5	که	۳۵۳۳۳	۳۶۰۰۰	۴.۵	۰.۷
6	است	۳۳۵۸۱	۳۰۰۰۰	۴.۵	۰.۸
7	این	۲۸۲۱۶	۲۵۷۱۴.۳	۴.۴	۰.۸
8	را	۲۴۲۶۳	۲۲۵۰۰	۴.۴	۰.۹
9	با	۲۲۹۲۱	۲۰۰۰۰	۴.۴	۰.۹
10	آن	۱۳۰۸۰	۱۸۰۰۰	۴.۱	۱



شکل ۱.نمودار زیف رتبه ۱ تا ۴۰ در پیکره ی مقالات و کتابها

جدول ۲. بسامد واژگان پر تکرار در پیکره‌ی کتابها

رتبه	انواع کلمه	بسامد واقعی	بسامد تخمینی	لگاریتم ۱۰ بسامد واقعی	لگاریتم ۱۰ رتبه
1	و	۱۱۷۲۷	۲۵۵۰۰	۴.۰۶	۰
2	به	۷۷۳۶	۱۲۷۵۰	۳.۹	۰.۳
3	که	۷۵۰۷	۸۵۰۰	۳.۹	۰.۵
4	از	۶۰۳۸	۶۳۷۵	۳.۸	۰.۶
5	را	۵۴۴۱	۵۱۰۰	۳.۷	۰.۷
6	در	۴۷۹۹	۴۲۵۰	۳.۷	۰.۸
7	بود	۳۷۰۶	۳۶۴۲.۸۵۷	۳.۶	۰.۸
8	با	۲۷۶۶	۳۱۸۷.۵	۳.۴	۰.۹
9	گفت	۲۵۹۳	۲۸۳۳.۳۳۳	۳.۴	۰.۹
10	این	۲۳۷۸	۲۵۵۰	۳.۴	۱

جدول‌های نشان داده شده رتبه ۱ تا ۴۰ داده‌ها محاسبه شده در دو پیکره‌ی مجموع سه کتاب داستان و کل مقالات را نشان می‌دهد. شکل ۱ نیز نمودار این دو پیکره در کنار هم را نشان می‌دهد. منحنی‌های رسم شده‌ی تخمینی بر مبنای داده‌ها و قانون زیف با منحنی‌های رسم شده براساس بسامد واقعی و رتبه تفاوت چندانی ندارد.

به منظور اعتبار سنجی کار نیز از ضریب همبستگی پیرسون استفاده شدو این ضریب همبستگی میان بسامد واقعی و تخمینی را می‌سنجد که نتایج آن در دو پیکره به ترتیب در پیکره مجموع مقالات ۰/۹۷ درصد و مجموع کتاب‌ها ۰/۹۴ محاسبه شد.

۳- نتایج

پژوهش‌های انجام شده در حوزه‌ی پردازش زبان طبیعی فارسی بسیار کم است. همچنین وجود نرم‌افزارهای پایه‌ای مناسب برای پردازش زبان فارسی آنقدر کم است که اساساً می‌توان گفت که هیچ نرم‌افزاری مناسبی وجود ندارد. این بحران را می‌شود با یک مثال توضیح داد. برای مثال برای ساخت خانه، به ابزار و مصالح خاص نیاز است تا بتوان خانه‌ای درخور و مناسب به عمل آورد. اگر پردازش زبان طبیعی فارسی، خانه محسوب شود. ابزار و مصالح مورد نیاز، داده‌ها و نرم‌افزارهای مناسب برای ساخته این خانه است .

پژوهش انجام شده در تلاش بود که با فراهم آوردن داه‌های به نسبت مناسب‌تر از داده‌های مورد تحلیل قرار گرفته در دیگر پژوهش‌های مشابه، نتایج به دست آمده از تحلیل آماری را بهبود بخشد. همینطور با اثبات تبعیت پراکندگی واژه‌های زبان فارسی از قانون زیف، که قانونی همگانی و جاری در تمام محیط‌های زبانی مختلف است، می‌توان در جهت سیستمی کردن زبان فارسی گام برداشت. از نتایج برآمده از داده‌های مورد تحلیل قرار گرفته می‌توان نتیجه گرفت که هر چه پیکره جمع‌آوری شده دارای پیش‌پردازش عمقی‌تر، منظم‌تر و بزرگتر باشد، نتایج تحلیل‌های آماری مختلف بر روی دادگان زبانی بهتر و سودمندتر خواهد بود و هر چه در جهت بهینه کردن این روند بیشتر تلاش شود، پردازش زبان طبیعی فارسی قدرتمندتر و سریعتر پیش‌خواهد رفت.

با استفاده از پیکره کل داده‌ها می‌توان به اشکالات موجود در نگارش و حتی رسم‌الخط فارسی پی برد و قانون‌هایی را وضع کرد که در جهت پیشبرد درونداد صحیح داده‌های زبان

فارسی مفید باشد. همچنین می‌توان الگوریتم‌های جامع‌تر و کامل‌تری برای طراحی نرم‌افزار پیش‌پردازش داده‌های فارسی کشید .

نکته مهم‌تر و قابل توجه‌تری که در این پژوهش نهفته است، رویکرد زبان‌شناسانه حاکم در تحلیل است. پژوهش‌های مشابه در این حوزه، هر کدام در حوزه‌های غیر از حوزه‌ی زبان‌شناسی و متناسب با نیاز همان حوزه‌ی خاص انجام شده بود.

از جمله کاربردهای علمی پژوهش استفاده از نتیجه اثبات قانون در بازیابی اطلاعات به زبان فارسی است. بازیابی اطلاعات داده‌های فارسی زبان نیز در سرعت و بهبودی پروژه‌های مختلف در پردازش زبان طبیعی فارسی است. همانطور که در قبل یز به این عناوین اشاره شده است، با استفاده از بازیابی اطلاعات به زبان فارسی می‌توان نرم‌افزارهای مورد نیاز در حوزه‌پردازش زبان طبیعی طراحی شود. از جمله نرم‌افزارهایی که می‌توان به آن اشاره کرد، نرم‌افزار استخراج چکیده از متن است، با توجه به میزان بسامد سطوح مختلف زبانی در متن می‌توان به محتوای مختلف موجود در متن دست پیدا کرد، مهمترین آنها را برگزید و چکیده‌ای جامع و کامل از متن ارائه داد. همچنین تحلیل آماری قانون بنیاد می‌تواند در استخراج الگوهای معتبرتری از پیکره‌ها استفاده شود و به اعتبار کار بیافزاید. با استفاده از الگوهای معتبر مستخرج از پیکره در حوزه‌های متفاوت، می‌توان نرم‌افزارهایی همچون تبدیل گفتار به نوشتار یا بالعکس، نرم‌افزارهای بن واژه ساز و ... گام‌های بزرگ و موثری برداشت.

منابع

[1] Kanwal, J., Smith, K. & Culbertson, J. (2017). Zipf's law of abbreviation and the principle of least effort: language users optimise a miniature lexicon for efficient communication. *Cognition*, 165, 45-52.

[2] Darrudi, E., Hejazi, M.R & Oroumchian, F. (2004). Assessment of a modern farsi Corpus. Retrieved June 8, 2017, from https://www.researchgate.net/publication/228605156

[3]Abdelali, A., Cowie, J. & Soliman, H. S. (2005). Building a modern standard arabic corpus. Retrieved January 1, 2005, from https://www.researchgate.net/publication/228958341

[4]Goweder, A. & De Roeck, A. (2001). Assessment of a significant

arabic corpus. Retrieved May 8, 2014, from https://www.researchgate.net/publication/233967788

[5]هاشم زاده، محمدجواد؛ نخعی، زینب؛ مرادی مقدم، حسین (۱۳۹۲) .کاربرد و تعدیل قانون زیف و الگوی بازو در بازشناسی واژه‌های بازدارنده زبان فارسی با استفاده از خوشه زبانی مقالات علمی - پژوهشی رشته کتابداری و اطلاع‌رسانی. پژوهش نامه کتابداری و اطلاع رسانی،۳ (۲)، ۲۰۸-۱۹۱.

[6]مهدی نسب، ت. (۱۳۹۲). *شناسایی عوامل مؤثر بر انتخاب منابع اطلاعاتی اعضای هیأت علمی دانشگاه بیرجند بر اساس اصل کمترین کوشش زیف*. بازیابی شده از ایرانداک. (۲۲۲۱۷۲).

[7]ترابی، م. (۱۳۸۹). *بررسی روش‌ها و معیارهای کاربرد پیکره‌ها در آموزش زبان، با توجه ویژه به زبان فارسی*. بازیابی شده از ایرانداک. (۱۵۱۹۸۵).

[8]غروی قوقانی، م. (۱۳۸۵). *تعیین واژگان پایه فارسی معیار گفتاری در بزرگسالان و تجزیه و تحلیل تواتر آنها براساس قانون زیف*. بازیابی شده از ایرانداک. (۱۰۴۸۲۶).

[9] Taghiyareh, F., Darrudi, E., Oroumchian, F. & Angoshtari, N. (2003). Compression of persian text for web-based applications, without explicit decompression. Retrieved March 16, 2014, from https://www.researchgate.net/publication/241775191.

[10]Mehri, A. & Jamaati, M. (2017). Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. *Physics Letters A*, 218, 1-8.

[11]Chao-Lin Liu, Shuhua Zhang, Yuanli Geng, Huei-ling Lai, and Hongsu Wang. Character distributions of classical Chinese literary texts: Zipf's law, genres, and epochs, Proceedings of the 2017 International Conference on Digital Humanities (DH 2017), 507– 511, Montréal, Québec, Canada, 8-11 August 2017.

[12]Bentz, C.; Ferrer-i-Cancho, R. Zipf's law of abbreviation as a language universal. In Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics, Leiden, The Netherlands, 26–30 October 2015; Bentz, C., Jäger, G., Yanovich, I., Eds.; University of Tübingen: Tübingen, Germany, 2016.

[13]Manning, C. D. & Schiitze, H. (2000). *Foundations of statistical*

natural language processing. Cambridge, Massachusetts London, England: The MIT Press

تأثیر بسط پرسش با شبکه‌های واژگانی بر میزان بازخوانی سامانه بازیابی اطلاعات قرآن کریم برای فارسی زبانان: WordNet یا BabelNet ؟

پگاه تاجر[*]، سید مصطفی فخراحمد[**]، زهرا خدادادی[***] و عبدالرسول جوکار[****]

چکیده

این مقاله ضمن ارائه چارچوب پیشنهادی سامانه بازیابی اطلاعات معنایی قرآن کریم برای کاربران فارسی زبان، به مطالعه تأثیر دو شبکه واژگانی "وردنت" و "بابل نت" در میزان بازخوانی سامانه پیشنهادی می‌پردازد. روش شناسی پژوهش حاضر از نوع طراحی است که در آن از رویکرد مطالعه تجربی چند گروهی با پس آزمون صرف استفاده شده است. در این راستا، مجموعه آزمونی توسط محققان ساخته شد و ۱۲ آزمایش با ۹۰ پرسش انجام گرفت. در هر بار آزمایش یک سطح معنایی بسط (مترادفی، هایپرنیمی، هیپونیمی و مجموعه ترادف این سه سطح) بر ورودی‌های چهار سامانه پیاده سازی شده، اعمال گردید و شاخص بازخوانی سامانه‌ها محاسبه گردید. تجزیه و تحلیل داده‌ها با آزمون کروسکال والیس نشان داد که شبکه‌های واژگانی در میزان بازخوانی سامانه پیشنهادی تأثیر گذار هستند و شبکه واژگانی "بابل نت فارسی" به صورت معناداری منجر به افزایش بازخوانی سامانه پیشنهادی می‌گردد. همچنین، سطح بسط پرسش با مترادف ها، نیز به طور معناداری بیشتر از دیگر سطوح منجر به افزایش بازخوانی سامانه می‌گردد.

واژه‌های کلیدی: شبکه‌های واژگانی، بسط پرسش، بازخوانی، بازیابی اطلاعات قرآن، بابل نت، وردنت.

۱- مقدمه

در رویکرد معنایی به بازیابی اطلاعات که جستجوی "مفاهیم" به جای "واژگان" را مورد

---

* دانشجوی دکتری علم اطلاعات و دانش شناسی، دانشگاه شیراز/ عضو هیأت علمی گروه علم اطلاعات و دانش شناسی، واحد مرودشت، دانشگاه آزاد اسلامی، مرودشت، ایران، ptajer@shirazu.ac.ir

** استادیار بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات، دانشگاه شیراز، fakhrahmad@shirazu.ac.ir

*** استادیار گروه ریاضی، واحد مرودشت، دانشگاه آزاد اسلامی، مرودشت، ایران zhrkhodadadi@gmail.com

**** استاد بخش علم اطلاعات و دانش شناسی، دانشگاه شیراز، ajowkar2003@yahoo.com

توجه قرار می‌دهد، راهکارهایی چون نمایه سازی معنایی پنهان[۱]، نمایه سازی معنایی صریح[۲] و بسط پرسش[۳] از طریق اصطلاحنامه‌ها و هستان شناسی ها[۴] ارائه شده است[۱]. به طور کلی، بسط پرسش از روش‌های سودمند تقویت عملکرد یک سامانه بازیابی اطلاعات می‌باشد. در رویکردهای سنتی بسط پرسش، پرسش اولیه با بهره گیری از رویکردهای مختلفی چون بازخورد ربط[۵] و شبه بازخورد ربط[۶] مجددا فرمولبندی می‌گردد و برای جستجو در اختیار سامانه قرار می‌گیرد[۲]. بسط پرسش کاربر به کمک روابط معنایی موجود در اصطلاحنامه‌ها و هستان شناسی‌ها رویکرد نسبتا جدیدتری است که انتظار می‌رود منجر به درک بهتر پرسش کاربر، توسط سامانه گردد. در این رویکرد، هستان شناسی ها[۷] را در حکم ستون فقرات سامانه‌های بازیابی اطلاعات معنایی می‌دانند[۳]. یکی از انواع هستان شناسی‌ها شبکه واژگانی[۸] نام دارد که با نام واژهستان شناسی[۹] و یا واژگان معنایی نیز شناخته می‌شود. شبکه‌های واژگانی عمومی هستند و به حوزه خاصی اختصاص ندارند. "وردنت"[۱۰]، یکی از مشهورترین شبکه‌های واژگانی در زبان انگلیسی است. "بابل نت[۱۱]" نیز یک شبکه واژگانی چند زبانه است که زبان فارسی را پوشش می‌دهد.

با مطرح شدن مباحث بازیابی اطلاعات معنایی، پژوهشگران سامانه‌های بازیابی اطلاعات قرآن کریم نیز به استفاده از معنا روی آوردند تا بتوانند عملکرد این نوع سامانه‌ها را بهبود بخشند و

---

1. Latent Semantic Indexing (LSI)
نمایه سازی معنایی پنهان، ایجاد فضای مفهومی مدارک و ترم ها از روی مجموعه مدارک موجود می باشد.
2. Explicit Semantic analysis (ESA)
نمایه سازی معنایی صریح، ایجاد فضای مفهومی مدارک و ترم ها از طریق مجموعه‌های خارجی مانند پیکره ویکی پدیا می‌باشد.
3. Query Expansion (QE)
4. Ontologies
5. Relevance feedback
6. Pseudo-relevance feedback
۷. هستان شناسی ذکر خصیصه‌های روشن و صوری مفاهیم به اشتراک گذاشته شده است . به عبارت دیگر در یک هستان شناسی ابتدا بر اساس اصطلاحات موجود در جهان مدل شده، مجموعه مفاهیم تعریف می شوند، سپس روابط بین مفاهیم اعم از طبقه ای و غیر طبقه ای به صورت صریح تعریف می شود. در مرحله آخر مجموعه اصول بدیهی استنتاج می شوند. لازم به ذکر است که صوری بودن در تعاریف هستان شناسی، به این مسأله بر می گردد که یک آنتولوژی باید قادر باشد به وسیله ماشین شناخته شده و قابل خواندن باشد.
8. Lexicon
9. Lexical ontology
10. WordNet
11.BabelNet

مشکلات بازیابی بر اساس واژگان را مرتفع سازند.

از اوایل قرن بیست و یکم تا کنون، پژوهش‌های زیادی، به منظور استخراج دانش قرآنی صورت گرفته است. پژوهش هایی که به استخراج دانش از متن قرآن کریم پرداخته اند معمولا از فنون پردازش زبان طبیعی مانند رفع ابهام معنایی[۴]، [۵]، [۶] و ریشه یابی کلمات قرآنی[۷]،[۸]،[۹] بهره گرفته اند و هستان شناسی قرآن و پیکره‌های تفسیری[۱۰]،[۱۱] را تولید کرده اند.

از طرف دیگر، پژوهش هایی نیز وجود دارند که بر طراحی سامانه‌های پرسش و پاسخ[۱۲]و بازیابی اطلاعات قرآن کریم[۱۳] تمرکز داشته اند. فنون مورد استفاده در سامانه‌های بازیابی اطلاعات قرآنی را می‌توان به دو دسته عمده فنون جستجوی "کلیدواژه محور" و "معنا محور" تقسیم کرد. در سامانه‌های "کلیدواژه محور" نتایج بر اساس حروف کلمات موجود در پرسش برگردانده می‌شوند. لازم به ذکر است که اکثر ابزارهای جستجوی قرآنی از این فن بهره می‌برند. این در حالی است که این رویکرد عیوب عمده‌ای مانند بازیابی تعدادی آیه غیر مرتبط یا پرسش و عدم بازیابی تعدادی از آیات مربوط دارد[۱۴].

در سامانه‌های "معنا محور" نتایج از طریق انطباق معنای بافتاری کلمات پرسش با متن قرآن کریم برگردانده می‌شوند. راهکارهای جستجوی معنایی قرآنی موجود عبارتند از روش‌های "هستان شناسی مدار"[۱۵]، بسط پرسش با مجموعه مترادف ها[۱۶] و بازیابی اطلاعات بین زبانی[۱۷]. در روش‌های "هستان شناسی مدار" انطباق مفاهیم پرسش کاربر با متن قرآن انجام می‌شود. به این منظور از هستان شناسی‌های قرآنی استفاده می‌شود. مشکل عمده روش‌های "هستان شناسی مدار" این است که در این روش‌ها اغلب از یک هستان شناسی قرآنی استفاده می‌شود که همه مفاهیم قرآن کریم را دربرندارد. زیرا همه هستان شناسی‌های قرآنی ساخته شده به یکدیگر نگاشت نشده اند. بنابراین، پژوهشگران از هستان شناسی هایی استفاده کرده اند که از جامعیت لازم برخوردار نیستند. البته، تلاش‌های اندکی به منظور نگاشت هستان شناسی‌های قرآنی و رفع این مشکل آغاز شده است[۱۴] اما، چالش همچنان باقی است. در روش بسط پرسش با مجموعه مترادف ها، ابتدا تمامی مترادف‌های کلمات پرسش با استفاده از "وردنت" استخراج می‌شود سپس، تمام آیات حاوی هر یک از مترادف‌های کلمات پرسش برگردانده می‌شود. در بازیابی اطلاعات بین زبانی هم، ابتدا کلمات پرسش ورودی به زبان دیگر

ترجمه می‌شوند و سپس، آیاتی کـه شـامل کلمـات موجـود در پرسـش ترجمـه شـده هسـتند برگردانده می‌شوند.

مرور ادبیات پژوهش نشان می‌دهد که آن دسته از سامانه‌های بازیابی اطلاعات قرآنـی کـه در کنار فنون پردازش طبیعی مانند رفع ابهام معنایی و ریشه یابی، از بسـط پرسـش بـا روابـط معنایی مفاهیم و واژه‌ها بهره می‌برند، از اثر بخشی بالاتری نسبت به سامانه هـایی کـه فقـط بـر جستجوهای کلیدواژه‌ای استوارند، برخوردار هستند. همچنین، با توجه به چـالش عـدم نگاشـت جامع هستان شناسی‌های قرآنی از یک طرف و اهمیت بسط معنایی پرسش کاربر در بازیـابی از طرف دیگر، جای بررسی کارایی شبکه‌های واژگانی موجود در زبان‌های مختلـف بـه منظـور بـه کارگیری در سامانه‌های بازیابی اطلاعات قرآنی در ادبیات پژوهش خالی است.

از سوی دیگر، اغلب پژوهش هایی که به طراحی سامانه‌های بازیابی اطلاعات معنایی قرآن کریم پرداخته اند، متن عربی و انگلیسی قرآن کریم را مورد توجه قرار داده اند. انـدک پـژوهش هایی نیز بر طراحی سامانه‌های قرآنی برای کاربران مالایی[۱۸]، [۱۹] و اندونزیایی[۲۰] تمرکـز کرده اند. به طور کلی، پژوهش‌های مبتنی بر متن ترجمه قرآن به زبان‌های دیگـر بسـیار انـدک هستند. همچنین، اکثر پژوهش‌های مبتنی بر بسط پرسش با شبکه‌هـای واژگـانی، از "وردنـت" بهره برداری کرده اند.

با توجه به این که، به بازیابی اطلاعـات معنـایی قـرآن کـریم بـرای فارسـی زبانـان کمتـر پرداخته شده است، این مقاله ضمن ارائه چارچوب پیشنهادی سامانه بازیابی اطلاعـات معنـایی قرآن کریم برای کاربران فارسی زبان، به مطالعـه تـأثیر دو شـبکه واژگـانی مشـهور و قدرتمنـد "وردنت" و "بابل نت" در میزان بازخوانی سامانه پیشنهادی می‌پردازد. لازم به ذکر است که بـا وجود این که پژوهش‌های زیادی بر سامانه‌های بازیابی اطلاعـات مبتنـی بـر "وردنـت" تمرکـز کرده اند، تاکنون میزان کارایی "وردنت" در مقابل "بابل نت" به بوته آزمـایش گذاشـته نشـده است.

در ادامه فرضیه‌های پژوهش و روش شناسی آورده می‌شوند. سپس یافته‌ها ارائه می‌گردنـد و در پایان، به بحث و نتیجه گیری پرداخته می‌شود.

۲- فرضیه‌های پژوهش

همانطور که از پیشینه پژوهش بر می‌آید، به طور کلی بسط پرسش اولیه کـاربر منجـر بـه

افزایش بازخوانی سامانه‌های بازیابی اطلاعات می‌گردد. اما این امر، در بازیابی اطلاعات بین زبانی کاملا روشن نیست. بر این اساس، فرضیه‌های این پژوهش عبارتند از:

۱- شبکه واژگانی "وردنت" برمیزان بازخوانی سامانه پیشنهادی بازیابی اطلاعات قرآن کریم برای فارسی زبانان تأثیر دارد.

۲- شبکه واژگانی "بابل نت انگلیسی" بر میزان بازخوانی سامانه پیشنهادی بازیابی اطلاعات قرآن کریم برای فارسی زبانان تأثیر دارد.

۳- شبکه واژگانی "بابل نت فارسی" منجر به افزایش بازخوانی سامانه پیشنهادی بازیابی اطلاعات قرآن کریم برای فارسی زبانان می‌گردد.

۳- روش شناسی

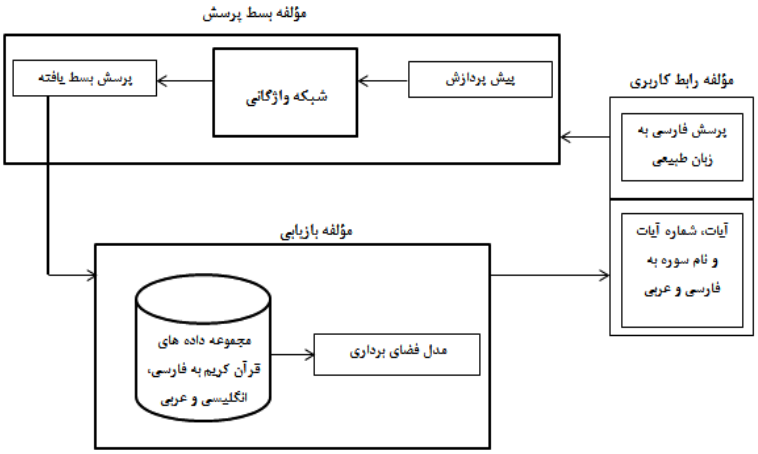
روش شناسی پژوهش حاضر، از نوع طراحی است که در آن از رویکرد مطالعه تجربی چند گروهی با پس آزمون صرف[1] استفاده شده است.

در ادامه، ضمن ارائه چارچوب مفهومی پیشنهادی سامانه بازیابی اطلاعات معنایی قرآن کریم برای کاربران فارسی زبان مراحل اجرای پژوهش تشریح می‌گردد:

۳-۱- چارچوب مفهومی پیشنهادی

چارچوب مفهومی پیشنهاد شده به منظور طراحی یک سامانه بازیابی اطلاعات معنایی قرآن کریم برای فارسی زبانان شامل سه مؤلفه رابط کاربری، بسط پرسش و بازیابی می‌باشد (شکل ۱). سامانه پیشنهادی ابتدا پرسش کاربر را به زبان فارسی دریافت می‌کند، با رویکرد کیسه کلمات[2] به بسط معنایی پرسش با استفاده از یک شبکه واژگانی می‌پردازد و فرآیند بازیابی آیات، شماره آنها و نام سوره را به دو زبان فارسی و عربی در مدل فضای برداری انجام می‌دهد.

1. Experimental Posttest-Only Design
2. Bag-of-words



شکل۱: چارچوب مفهومی پیشنهادی بازیابی اطلاعات معنایی قرآن کریم برای کاربران فارسی زبان

۳-۲- مراحل اجرای پژوهش

الف: ساخت داده‌های آزمون: با توجه به عدم دسترسی به مجموعه آزمون[1] مناسب به منظور محاسبه شاخص بازخوانی[2] سامانه، مجموعه داده‌های آزمون این پژوهش توسط پژوهشگران ایجاد شد. برای ساخت مجموعه داده‌های آزمون مناسب این پژوهش لازم بود تا به متخصصان موضوعی قرآن کریم و خبرگان علم تفسیر مراجعه شود تا به قضاوت ربط پرسش‌ها با آیات قرآنی بپردازند. خوشبختانه منابع معتبری چون تفاسیر موضوعی قرآن کریم در دسترس می‌باشند که حاوی قضاوت ربط متخصص برای تعین دسته بندی‌های موضوعی آیات قرآن کریم می‌باشند. از این روی، این منابع اساس ساخت مجموعه داده‌های آزمون این پژوهش قرار گرفتند. وب سایت دانشنامه موضوعی قرآنی تبیان[3] در دسترس می‌باشد که براساس تفاسیر معتبر موضوعی قرآن کریم مانند تفسیر المیزان ایجاد شده است. انتخاب دانشنامه موضوعی قرآنی تبیان به توصیه خبرگان علوم قرآنی صورت گرفت. ۹۰ موضوع از این دانشنامه به صورت تصادفی به همراه شماره‌های آیات و سوره‌های مربوط به آنها استخراج شد و به عنوان

1. Test Collection
2. Recall
3. www.tebyan.net

پرسش‌های داده‌های آزمون این پژوهش ذخیره شدند.

۱- ب: پیاده سازی سامانه: متون قرآنی استفاده شده در پیاده سازی سامانه عبارتند از ترجمهٔ آیت‌الله مکارم شیرازی از قرآن کریم به فارسی، ترجمهٔ "آربری"[1] از قرآن به انگلیسی و متن عربی قرآن کریم از وبگاه تنزیل[2]. برای پیاده‌سازی سامانه نیز از زبان برنامه‌نویسی پایتون (نسخهٔ ۲,۷,۱۲) استفاده شد. مراحل اجرایی پیاده سازی سامانه به شرح زیر است:

۲- ۱- پیش‌پردازش: پیش‌پردازش متون هم برای پرسش‌ها و هم برای متن ترجمهٔ قرآن به فارسی با استفاده از کتابخانهٔ هضم (نسخهٔ ۰,۵,۲) صورت گرفت. در این راستا مراحل زیر انجام شد:

- نرمال‌سازی: جایگزینی «ی» و «ک» عربی با فارسی، اتصال «می» و «نمی» به واژهٔ بعد از خود با نیم‌فاصله، اتصال «تر»، «ترین» و «ها»ی جمع به واژهٔ پیش از خود با نیم‌فاصله و ... .

- ریشه‌یابی: حذف وندهای تصریفی، ارائهٔ ریشهٔ افعال و ... . لازم به ذکر است که در حالت‌هایی که بسط پرسش با استفاده از منابع انگلیسی انجام شده است، از بستهٔ "ان. ال. تی. کی. .استم"[3] برای ریشه‌یابی استفاده شده است.

- قطعه‌بندی: تشخیص مرز واژه‌ها

- حذف ایست واژه ها[4]: وجود برخی واژه‌ها (مانند «و»، «از»، «با»، ...) در پرسش باعث پایین آمدن دقت جستجو می‌شود (برای مثال سیستم تمام آیات دارای واژهٔ «و» را بر می‌گرداند!). لذا این واژه‌ها از پرسش حذف شدند.

۲- بسط پرسش: بسط پرسش با استفاده از دو دادهٔ وردنت (نسخهٔ ۳,۰) و بابل‌نت (نسخهٔ ۳,۷) انجام شد.

برای بسط پرسش با وردنت، ابتدا با استفاده از دادهٔ متنی یک واژه‌نامهٔ فارسی به انگلیسی معادل انگلیسی برای واژه‌های پرسش پیدا شد. واژه‌نامه‌ها معمولاً برای هر واژه چند معادل پیشنهاد می‌کنند. در این پژوهش از اولین معادل برای هر واژه استفاده شد.

پس از مشخص شدن معادل هر واژه، جستجو بر اساس آن در وردنت انجام گرفت. برای

---

1. Arberry
2. tanzil.net
3. nltk.stem
4.Stopwords

به دست آوردن هایپرنیم ها[1]، در ذیل هر یک از مجموعه‌های ترادف ارائه‌شده، مواردی که به عنوان هایپرنیم مستقیم ارائه شده‌اند در نظر گرفته شد. روال کار برای هیپونیم ها [2]هم به همین منوال بود و هیپونیم‌های مستقیم مورد استفاده قرار گرفت. نهایتا فهرست بسط‌یافتهٔ واژه‌ها پس از حذف موارد تکراری و ریشه‌یابی به جستجوگر داده شد.

بسط پرسش با استفاده از دادهٔ بابل‌نت به صورت مجزا یک بار با بخش فارسی بابل‌نت و یک بار با بخش انگلیسی بابل‌نت انجام شد. برای استفاده از بخش انگلیسی بابل‌نت مانند وردنت از یک واژه‌نامهٔ فارسی به انگلیسی استفاده شد. پیاده‌سازی بسط پرسش با بخش انگلیسی بابل‌نت به منظور مقایسهٔ بهتر با وردنت صورت گرفت.

۳- مدل فضای برداری[3]: مدل فضای برداری از مدل‌های بسیار مشهور و پر استفاده بازیابی اطلاعات است که در آن، پرسش‌ها و مدارک به صورت بردارهایی از کلمات بازنمون می‌شوند و برای هر کلمه وزنی اختصاص داده می‌شود. به صورت زیر:

$$\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$

$$\vec{dj} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

سپس میزان شباهت بردارها محاسبه می‌گردد. مدل فضای برداری بر مبنای اندازه‌گیری کسینوسی پیاده‌سازی شد و وزن‌دهی اصطلاحات بر اساس TF-IDF صورت گرفت. به صورت زیر:

- محاسبه TFبا فرمول (۱)

$$1 + \log10(\text{TermFreq}) \text{ if TermFreq} > 0 \quad (۱)$$
$$\text{if TermFreq} == 0$$

TermFreq = فرکانس واژه در آیه و یا پرسش

- محاسبه IDF با فرمول (۲)

$$\log10(\text{all\_aye\_count} / \text{aye\_freq}) \quad (۲)$$

فرمول (۲) برای محاسبه IDF یک واژه استفاده شده که all_aye_count تعداد کل آیات

---

1 . Hypernyms
2. Hyponyms
3. Vector Space Model (VSM)

می‌باشد و aye_freq تعداد آیه‌هایی است که این کلمه در آن‌ها آمده است.

- محاسبه TF-IDFبا فرمول (۳)

TF * IDF (۳)

شباهت کسینوسی نیز با استفاده از فرمول (۴) محاسبه گردید.

$$sim(d_j, q) = \frac{\sum_{i-1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}} \quad (۴)$$

۳- ۴- محاسبه شاخص بازخوانی: برای محاسبه این شاخص از فرمول (۵) استفاده شد.

۴- ۵ تعداد کل مدارک مربوط /تعداد مدارک مربوط بازیابی شده= بازخوانی

ج: آزمایش: به منظور بررسی فرضیه‌های پژوهش، ۱۲ آزمایش انجام گرفت. به عبارت دیگر، یک سامانه پایه[1] بدون شبکه واژگانی و سه سامانه آزمایشی مبتنی بر "وردنت"، مبتنی بر "بابل نت انگلیسی" و مبتنی بر "بابل نت فارسی" پیاده سازی شدند. در این پژوهش، مقایسه سامانه‌های آزمایشی در چهار سطح بسط یعنی روابط معنایی مترادفی[2]، هایپرنیمی[3] و هیپونیمی[4] و مجموعه ترادف[5] این سه نوع رابطه صورت پذیرفت. در هر بار آزمایش یک سطح معنایی بسط (متغیر مستقل)، بر تک تک ورودی‌های سامانه‌های آزمایشی (۹۰ پرسش) اعمال

---

1. Baseline
2. Synonymy
3.Hypernymy
4. Hyponymy
5. Synset

اطلاعات موجود در شبکه‌های واژگانی بر اساس یک تقسیم بندی معنایی به نام مجموعه ترادف مرتب گردیده اند که شامل لیستی از لغات مترادف می باشد که توسط یک اشاره گر معنایی با مجموعه ترادف دیگر در ارتباط است. این اشاره گر در واقع روابط معنایی خاص را نشان می دهد. برخی از روابط معنایی یک مفهوم عبارتند از رابطه هایی از نوع مترادف، متضاد، ابرمفهوم، زیرمفهوم(IS-A) ،جزئیات(Part of) ، شمول (Has-A) و ... . در این پژوهش مجموعه ترادف عبارتند از، مترادف ها، هایپرنیم ها و هیپونیم‌های یک مفهوم در شبکه واژگانی "وردنت" و "بابل نت".

شد. پس از طی شدن فرآیند بازیابی اطلاعات در هر سامانه و بررسی خروجی آن، شاخص بازخوانی سامانه (متغیر وابسته) محاسبه و ثبت گردید و مبنای مقایسه قرار گرفت. لازم به ذکر است که در ادبیات بازیابی اطلاعات، ارزیابی مطلوب سامانه‌ها حداقل با ۳۰ پرسش صورت می‌گیرد.

۵- **د: روش تجزیه و تحلیل داده ها:** تجزیه و تحلیل داده‌ها با استفاده از نرم افزار علوم اجتماعی[1] (نسخه ۲۳) صورت پذیرفت. با توجه به معناداری آزمون کولموگروف – اسمیرنف ($p<۰/۰۱$)، نمره‌های متغیر وابسته دارای توزیع نرمال نبود لذا برای مقایسه میانگین‌های نمرات بازخوانی سامانه‌های مورد مطالعه از معادل ناپارامتری تحلیل واریانس یکطرفه یعنی کروسکال والیس[2] استفاده شد.

۴- یافته ها

نتایج آزمون کروسکال والیس نشان داد که شبکه واژگانی "بابل نت فارسی" با میانگین رتبه‌ای ۱۱۱۷,۲۵باعث افزایش بازخوانی سامانه پیشنهادی می‌گردد. این در حالی است که بسط پرسش با "وردنت" و "بابل نت انگلیسی" میزان بازخوانی سامانه را کاهش می‌دهد. این نتایج با توجه به میزان خی دو به دست آمده از این آزمون (x^2=۳۰۹/۷۶۷) در سطح ۹۹ درصد معنی دار است (جدول۱)

جدول۱: آزمون معنی داری تفاوت میانگین نمره بازخوانی سامانه ها

سامانه	میانگین رتبه ای	x^2	Df	Sig
سامانه پایه	۱۰۲۲,۴۱	۳۰۹,۷۶۷	۴	./۰۰۰
سامانه مبتنی بر "وردنت"	۶۳۱,۹۶			
سامانه مبتنی بر "بابل نت انگلیسی"	۶۷۲,۹۱			
سامانه مبتنی بر "بابل نت فارسی"	۱۱۱۷,۲۵			

نتایج آزمون کروسکال والیس حاکی از این است که "بابل نت فارسی" نسبت به "وردنت"

1.SPSS
2. Kruskal Wallis Test

و "بابل نت انگلیسی" در سطوح مختلف بسط پرسش (بسط با مترادف ها، هایپر نیم ها، هیپونیم‌ها و مجموعه ترادف این سه سطح) منجر به افزایش بازخوانی سامانه پیشنهادی می‌گردد ($p<۰/۰۱$) (جداول ۵-۲).

جدول۲: آزمون معنی داری تفاوت میانگین نمره بازخوانی سامانه‌ها در سطح بسط پرسش با مترادف ها

سامانه	میانگین رتبه ای	x^2	Df	Sig
سامانه پایه	۲۶۰/۸۸	۹۷/۳۱	۴	۰/۰۰۰
سامانه مبتنی بر "وردنت"	۱۵۰/۹۹			
سامانه مبتنی بر "بابل نت انگلیسی"	۱۶۱/۰۳			
سامانه مبتنی بر "بابل نت فارسی"	۲۸۰/۹۶			

جدول۳: آزمون معنی داری تفاوت میانگین نمره بازخوانی سامانه‌ها در سطح بسط پرسش با هایپرنیم ها

سامانه	میانگین رتبه ای	x^2	Df	Sig
سامانه پایه	۲۵۷/۲۴	۷۵/۸۵	۴	۰/۰۰۰
سامانه مبتنی بر "وردنت"	۱۵۹/۶۲			
سامانه مبتنی بر "بابل نت انگلیسی"	۱۶۸/۱۲			
سامانه مبتنی بر "بابل نت فارسی"	۲۷۷/۸۹			

جدول۴: آزمون معنی داری تفاوت میانگین نمره بازخوانی سامانه‌ها در سطح بسط پرسش با هیپونیم ها

سامانه	میانگین رتبه ای	x^2	Df	Sig
سامانه پایه	۲۶۳/۷۸	۸۷/۰۵	۴	۰/۰۰۰
سامانه مبتنی بر "وردنت"	۱۵۶/۱۶			
سامانه مبتنی بر "بابل نت انگلیسی"	۱۶۳/۲۷			

سامانه مبتنی بر "بابل نت فارسی"	۲۸۰/۴۸			

جدول۵: آزمون معنی داری تفاوت میانگین نمره بازخوانی سامانه‌ها در سطح بسط پرسش با مجموعه ترادف سه سطح مترادفی، هایپرنیمی و هیپونیمی

سامانه	میانگین رتبه ای	x^2	Df	Sig
سامانه پایه	۲۴۲	۵۵/۹۶	۴	۰/۰۰۰
سامانه مبتنی بر "وردنت"	۱۶۶/۴۴			
سامانه مبتنی بر "بابل نت انگلیسی"	۱۸۱/۴۹			
سامانه مبتنی بر "بابل نت فارسی"	۲۷۹/۹۸			

با توجه به جداول ۲ الی ۵ می‌توان دریافت که بیشترین بازخوانی سامانه مبتنی بر "بابل نت فارسی" زمانی به بار می‌آید که بسط پرسش در سطح مترادفی انجام شود. لازم به ذکر است که سطح بسط هایپرنیمی منجر به کمترین بازخوانی در این سامانه می‌گردد.

به منظور بررسی دقیق تر معناداری تفاوت میانگین نمره بازخوانی سامانه‌های پیاده سازی شده، آزمون مقایسه‌های زوجی[1] کروسکال والیس نیز انجام شد. نتایج این آزمون نشان داد که تنها تفاوت میانگین نمره بازخوانی دو سامانه مبتنی بر "وردنت" و "بابل نت انگلیسی" معنادار نیست و بقیه مقایسه‌های زوجی در سطح ۹۹ درصد معنادار می‌باشند (جدول۶).

1. Pairwise Comparisons

جدول۶: آزمون معناداری مقایسه‌های زوجی میانگین نمره بازخوانی سامانه ها

	سامانه پایه	سامانه مبتنی بر "وردنت"	سامانه مبتنی بر "بابل نت انگلیسی"	سامانه مبتنی بر "بابل نت فارسی
سامانه پایه		۰/۰۰۰	۰/۰۰۰	۰/۰۰۰
سامانه مبتنی بر "وردنت"	۳۹۰/۴۵		*۱,۰۰۰	۰/۰۰۰
سامانه مبتنی بر "بابل نت انگلیسی"	۳۴۹/۵۰	-۴۰/۹۵		۰/۰۰۰
سامانه مبتنی بر "بابل نت فارسی"	-۹۴/۸۴	-۴۸۵/۳۰	-۴۴۴/۳۴	

ارقام زیر قطر، آماره آزمون و ارقام بالای قطر، سطح معناداری است.

*= عدم معناداری

۵- نتیجه گیری

این پژوهش، یک مطالعه تجربی در ارتباط با تأثیر بسط پرسش با شبکه‌های واژگانی "وردنت" و "بابل نت" بر میزان بازخوانی سامانه پیشنهادی بازیابی اطلاعات قرآن کریم برای فارسی زبانان ارائه داد.

بر اساس یافته ها، می‌توان به این نتیجه رسید که شبکه‌های واژگانی در میزان بازخوانی سامانه پیشنهادی تأثیر گذار هستند و شبکه واژگانی "بابل نت فارسی" منجر به افزایش بازخوانی این سامانه می‌گردد. بنابراین، فرضیه‌های این پژوهش تأیید می‌شوند.

از طرف دیگر، با توجه به یافته‌های به دست آمده در رابطه با سطوح بسط در این پژوهش می‌توان اظهار کرد که نتایج این پژوهش در خصوص کارایی سطح بسط پرسش با مترادف‌ها با پژوهش شعیب و همکاران (۲۰۰۹)[۱۶] همسو است.

با توجه به عدم نگاشت هستان شناسی‌های موجود قرآن مجید و به عبارت دیگر، نبود هستان شناسی جامع قرآن، بهره گیری از "بابل نت فارسی" در سامانه‌های قرآنی بین زبانی فارسی و عربی می‌تواند راهگشا باشد.

کاهش بازخوانی سامانه‌های پیاده سازی شده با شبکه‌های واژگانی "وردنت" و "بابل نت انگلیسی" را می‌توان به فرآیند ترجمه پرسش کاربر از فارسی به انگلیسی نسبت داد که البته روشن شدن این امر به مطالعه بیشتری نیاز دارد. از آنجایی که شبکه‌های واژگانی انگلیسی مانند "وردنت" از نظر روابط معنایی واژه‌ها بسیار غنی هستند، به نظر می‌رسد با مجهز کردن سامانه به الگوریتم‌های ترجمه ماشینی مناسب، بتوان با اعمال شبکه‌های واژگانی انگلیسی نیز بازخوانی سامانه پیشنهادی را بهبود بخشید.

با توجه به نقش مهم هستان شناسی‌ها در بازیابی اطلاعات معنایی، پیشنهاد می‌شود مطالعاتی در رابطه با تأثیر ترکیب شبکه‌های واژگانی با هستان شناسی‌های موجود قرآنی بر میزان بازخوانی سامانه هایی از این دست انجام شود. به عبارت دیگر، نقش مکملی هستان شناسی‌ها برای شبکه‌های واژگانی مورد بررسی قرار گیرد.

منابع

[۱] Müller, C. and Gurevych, I., 2008. "Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval". in *CLEF*.. Springer.

[۲] Abberley, D., et al., 1999. "The THISL broadcast news retrieval system". in *ESCA Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio*..

[۳] Taye, M.M., 2010. "Understanding semantic web and ontologies: Theory and applications". arXiv preprint arXiv:1006.4567..

[۴] Tiun, S., et al., *2013. "*Word Sense Disambiguation for English Quranic IR System. in Advances in Information Technology for the Holy Quran and Its Sciences (32519)", *Taibah University International Conference on*. IEEE.

[۵] Mussa, S.A.-A. and Tiun, S., 2015. "Word Sense Disambiguation on English Translation of Holy Quran". *Bulletin of Electrical Engineering and Informatics,*. 4(3): p. 241-247.

[۶] Mohamed, O.J. and Tiun, S., 2015. "Word sense disambiguation

based on yarowsky approach in english quranic information retrieval system". *Journal of Theoretical and Applied Information Technology*,. 82(1): p. 163..

[۷] Al-Taani, A.T. and Al-Gharaibeh, A.M., 2011. "Searching Concepts and Keywords in the Holy Quran".

[۸] Aljaloud, H., Dahab, M., and Kamal, M., 2016. "Stemmer impact on Quranic mobile information retrieval performance". *Int. J. Adv. Comput. Sci*. Appl.(IJACSA),. **7**(1).2, p. 135-139.

[۹] Al Gharaibeh, A., Al-TaanAl-Taan, A. i, and Alsmadi, I., 2011. "The usage of formal methods in Quran search system". in *Proceedings of international conference on information and communication systems, Ibrid, Jordan*..

[۱۰] Khan, H.U., et al., 2013. "Ontology based semantic search in Holy Quran". *International Journal of Future Computer and Communication*,. 2(6): p. 570.

[۱۱] AlMaayah, M., Sawalha , M., and Abushariah , M., *2014*. "A proposed model for Quranic Arabic WordNet". in *Proceedings of the 2nd Workshop on Language Resources and Evaluation for Religious Texts, 31 May, Reykjavik, Iceland*. LRA.

[۱۲] Abdelnasser, H., et al., 2014. "Al-Bayan: an arabic question answering system for the holy quran". in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*..

[۱۳] Othman, R. and Wahid , F.A., 2014. "Quranic texts retrieval in Indri". in *Information and Communication Technology for The Muslim World (ICT4M), The 5th International Conference on*. IEEE..

[۱۴] Alqahtani, M. and Atwell, E. A., 2015. "Review of Semantic Search Methods to Retrieve Information from the Qur'an Corpus".

[۱۵] Yauri, A.R., et al., 2013. "Quranic verse extraction base on concepts using OWL-DL ontology". *Research Journal of Applied Sciences, Engineering and Technology*,. 6(23), . p. 4492-4498

[۱۶] Shoaib, M., et al., 2009. "Relational WordNet model for semantic search in Holy Quran". in *Emerging Technologies,. ICET 2009. International Conference on*. IEEE.

[۱۷] Yunus, M., Zainuddin, R. , and Abdullah, N., 2010. "Semantic query for Quran documets results" . in *Open Systems (ICOS), IEEE Conference on*. IEEE..

[۱۸] Yunus, M.A., Mustapha, M., , A., and Samsudin, N.A., 2017. "Analysis of translated query in Quranic Malay and English translation documents with stemmer". in *MATEC Web of Conferences* EDP Sciences

[۱۹] Ahmad, N.D., Bennett, B., and Atwell, E., 2016. "Semantic-based Ontology for Malay Quran Reader". in *IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies*.

[20] Putra, S.J., et al., 2016. "A semantic-based question answering system for indonesian translation of Quran". in *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services. ACM.*

برچسب‌زنی اجزای سخن در نوشته‌های فارسی با استفاده از بازنمایی کلمات و شبکه عصبی بازگشتی RNN

عرفان رحمانی* و **سیامک سرمدی****

چکیده

مشخص کردن نقش دستوری کلمات در جمله (به عنوان مثال فعل، اسم و مفعول) با استفاده از برچسب‌هایی که به کلمات زده می‌شود را برچسب‌زنی اجزای سخن می‌گویند. برچسب‌زنی یکی از ابزارهای میانی مهم برای انجام اعمال دیگر مانند تجزیه و تحلیل دستوری و ترجمه ماشینی است. تحقیقات زیادی در این زمینه انجام شده‌است ولی با توجه به تفاوت قوانین مورفولوژیکی زبان ها، جداکردن کلمات و نوع برچسب‌ها درهر زبان متفاوت می‌باشد. برای رسیدن به دقت بیشتر در این زمینه از متدهای مختلفی (از جمله مدلهای زبانی و آماری) استفاده شده است. این مقاله به بررسی برچسب‌زنی اجزای سخن در زبان فارسی، با استفاده از شبکه‌های عصبی بازگشتی RNN می‌پردازد و یک مدل RNN که به دقت ۹۷/۳۶ درصد دست یافته است را معرفی می‌نماید.

واژه‌های کلیدی: برچسب زنی اجزای سخن، شبکه عصبی بازگشتی، بازنمایی کلمات

۱. مقدمه

توانایی شبکه‌های عصبی مصنوعی در یادگیری موازی از یک طرف و اهمیت یادگیری در پردازش زبان از طرف دیگر باعث شده که محققان به این روش پردازش علاقه پیدا کنند. محققان برای برچسب زنی کلمات جمله از شبکه‌های عصبی به نسبت کمتر استفاده کرده‌اند. پارامترهای زیادی روی کارایی برچسب‌زنی تاثیر می‌گذارند. اندازه‌ی داده‌های آموزشی، مکانیزم برچسب‌زنی، نوع زبان، و حتی روش‌های پیاده‌سازی همگی پارامترهای هستند که روی کارایی برچسب‌زنی تاثیر دارند. استفاده از شبکه عصبی برای برچسب‌زنی، یک موضوع جدید تحقیق

* دانشگاه صنعتی ارومیه، rahmani.erfan.71@gmail.com

** دانشگاه صنعتی ارومیه، siamaksarmady@uut.ac.ir

در نظر گرفته نمی شود، اما اخیرا توجه زیادی به این موضوع می‌شود. بخصوص شبکه‌های عصبی عمیق که در چند سال اخیر ابداع و مورد استفاده قرار گرفته اند، نتایج خوبی را در بسیاری از حوزه‌ها نشان داده اند. توان بالای یادگیری این شبکه‌ها در زمینه پردازش زبان نیز مورد استفاده قرار گرفته است.

در این تحقیق ما از شبکه عصبی عمیق RNN به عنوان ابزار اصلی پردازش متن استفاده می‌کنیم. این شبکه‌ها به مجموعه‌ای از پارامترها بستگی دارند که مقادیر آنها ممکن است در پیاده‌سازی تاثیرات مهمی داشته باشند. بنابراین برای بدست آوردن نتایج مناسب، پارامترهای برچسب‌زنی باید با دقت انتخاب شوند. همچنین طراحی شبکه عصبی نیز باید به نحوی باشد که برای برچسب‌زنی مناسب باشد.

۲ پیشینه تحقیق

تعدادی از برچسب‌زن‌های موجود زبان فارسی ابتدا نوشته‌های فارسی را به فرم لاتین آن تغییر می‌دهند. این برچسب‌زن‌ها از مجموعه نوشته بی‌جن‌خان و یا مجموعه‌های کاهش یافته از آن مجموعه نوشته استفاده کرده‌اند. برای پرهیز از مشکلاتی جزئی که مجموعه نوشته بی‌جن‌خان دارد، در این تحقیق از مجموعه نوشته مناسب‌تری به نام UPEC استفاده می‌شود.
در مقاله مژگان سراجی و همکاران [۱]، ابتدا نوشته فارسی پیش پردازش می‌شود و فاصله‌ها و نیم فاصله‌ها و نوشتارهای مختلف به یک حالت مشخص تبدیل می‌شوند، سپس جمله‌ها و کلمات جدا می‌شوند. بعد از آن، برچسب‌زنی اجزای سخن انجام می‌شود و در مرحله‌ی آخر نیز تجزیه جملات انجام می‌شود. روش برچسب زنی این سیستم از مقاله [۲] گرفته شده که در آن سیستم HunPOS برای زبان فارسی پیاده شده‌است. HunPOS یک پیاده سازی مجدد از برچسب‌زنی برپایه HMM (trigram) است که به کاربر اجازه می‌دهد برچسب‌زنی را با ویژگی‌های مختلفی انجام دهد. این سیستم علاوه بر سرعت و دقت بیشتر، الگوریتم تجزیه و تحلیل مورفولوژی را هم بکار گرفته‌است و به صورت متن باز در دسترس است. دقت برچسب-زنی در این سیستم به ۹۷ درصد رسیده‌است [۲].
شبکه‌های عصبی عمیق از نظر بازده قدرتمند بوده و توان انجام محاسبات موازی برای مراحل مختلف عملیاتشان را دارند. شبکه‌های عصبی اگرچه جزو مدل‌های آماری به حساب می‌آیند،

توان یادگیری محاسبات پیچیده را دارند. بنابراین روشن است که یادگیری یک روش مستقل از دامنه برای مسائل دنباله به دنباله مفید است. تحقیقات زیادی بر روی یادگیری دنباله به دنباله با شبکه‌های عصبی، انجام گرفته‌است که تعدادی از آنها معرفی خواهد شد.

شبکه‌های عصبی بازگشتی RNN مشهورترین شبکه‌هایی هستند که روی داده‌های متوالی به کار گرفته شده اند. آنها یک دنباله بردار را به عنوان ورودی گرفته (x1,x2,...,xn) و یک دنباله دیگر را در خروجی پس می‌دهند (h1,h2,…,hn). این شبکه‌ها در هر مرحله مقداری اطلاعات را از دنباله ورودی بازنمایی می‌کنند. اگرچه در تئوری شبکه‌های RNN می‌توانند وابستگی‌های طولانی را یاد بگیرند، در عمل برای انجام این کار موفق نبوده‌اند و بیشتر توان یادگیری آخرین ورودی‌ها (دنباله‌های کوتاه) را دارند [۳].

شبکه‌های LSTM برای مقابله با این مشکل طراحی شده‌اند. این شبکه‌ها از یک سلول حافظه استفاده کرده‌اند و نشان داده‌اند که وابستگی‌ها را در فاصله‌های زمانی طولانی حفظ می‌کنند. این شبکه‌ها برای اولین بار در مقاله Sepp Hechreiter [۴] معرفی شدند.

در مقاله Kyunghyun Cho [۵] یک مدل جدید شبکه عصبی معرفی شده‌است که شبکه عصبی بازگشتی رمزگذار و رمزگشا نامیده می‌شود. این مقاله از این شبکه‌های عصبی برای ترجمه ماشینی آماری استفاده کرده‌است. این مدل شامل دو شبکه عصبی RNN است. یک شبکه RNN که دنباله کلمات ورودی را به بردارهایی با طول ثابت تبدیل می‌کند. یک RNN دیگر که این بردارها را به دنباله کلمه خروجی تبدیل می‌کند. در این مدل نشان داده شده که بازنمایی‌های معنایی و دستوری مناسبی برای عبارات یاد گرفته شده‌اند.

مقاله Nal Kalchbrenner و همکاران [۶] علاوه بر تولید بازنمایی‌های پیوسته برای کلمات، عبارات و جملات ترجمه متن نیز انجام می‌دهد و تنها بر ترجمه بخش‌هایی از متن تکیه نکرده است. مقاله Alex Graves [۷] یک مکانیزم رسیدگی متفاوت و جدید را معرفی کرده است که به شبکه‌های عصبی اجازه می‌دهد روی قسمت‌های خاصی از ورودی خود تمرکز کنند. همین ایده به طور موفق و کارآمدی در مقاله Dzmitry Bahdanau[۸] بر روی یک مدل ترجمه ماشینی اعمال شده است. این مکانیزم رسیدگی، روی قسمت‌هایی از جمله ورودی که در تولید جمله خروجی تاثیر بیشتری دارند تمرکز می‌کند.

مقاله Ilya Sutskever [۹] با مدلی شبیه به مقاله [۵] و با استفاده از روش بازنمایی

کلمات و عبارات و جملات که در مقاله [۶] مطرح شد، جملات ورودی را به صورت معکوس می‌خواند. معکوس کردن جمله ورودی کیفیت LSTM را به طور قابل توجهی بهبود داده است، زیرا این کار بسیاری از وابستگی‌های کوتاه مدت بین جمله ورودی و جمله خروجی را نشان می‌دهد (دنباله خروجی به آخرین ورودی‌ها وابستگی نزدیک‌تری دارد).

مقاله‌های [۵] و [۸] در ترجمه جملات طولانی نسبت به مقاله [۹] ضعیف‌تر عمل کرده‌اند. در مقاله Zhiheng Huang [۱۰] مدل‌هایی با استفاده از LSTM برای کارهای متوالی معرفی شده‌اند. این مقاله برای اولین بار شبکه عصبی دو طرفه LSTM-CRF را برای کارهای متوالی در پردازش زبان طبیعی به کار برده‌است و به این نتیجه رسیده است که مدل BI-LSTM-CRF بهترین کارایی را از بین دیگر مدل‌ها دارد.

در مقاله [۱۱] یک معماری شبکه عصبی جدید معرفی شده است. در این معماری از هر دو بازنمایی سطح کاراکتر و بازنمایی سطح کلمه در یک معماری ترکیبی از LSTM، CNN و CRF استفاده شده است. این سیستم انتها به انتها است و هیچ نیازی به مهندسی ویژگی‌ها و آماده سازی داده‌ها ندارد. این سیستم روی دو عمل برچسب زنی متوالی (برچسب زنی اجزای سخن و NER در زبان انگلیسی) آزمایش شده است و در برچسب زنی اجزای سخن به دقت ۹۷ و در NER به دقت ۹۱ درصد دست یافته است.

در مقاله [۱۲] که یک مدل شبیه به مدل مقاله [۱۱] است یک شبکه عصبی LSTM دو طرفه با استفاده از CRF معرفی شده است. این مدل بازنمایی سطح کاراکتر کلمات و بازنمایی کلمات با استفاده از روش ارائه شده توسط Mikolov [۱۳] را با هم ترکیب کرده است.

۱-۲ مجموعه نوشته UPEC

مژگان سراجی و همکاران [۱] تصمیم به دستکاری، تعمیر و نرمال‌سازی مجموعه نوشته بی‌جن‌خان [۱۴] گرفتند و مجموعه نوشته‌ای به نام UPEC را ساختند. در این مجموعه نوشته تمامی قوانین مربوط به فاصله و نیم فاصله که در زبان فارسی مطرح است رعایت شده‌است همچنین این مجموعه شامل یک مجموعه برچسب با ۳۱ برچسب است. در این مقاله از این مجموعه نوشته برای آموزش و آزمایش شبکه عصبی RNN استفاده شده است.

۲-۲ تعبیه نمودن کلمات فارسی

در شبکه‌های عصبی معمولا در دنباله کلمه ورودی، هر کلمه را با یک بردار نمایش میدهند. بهترین روش‌های تبدیل کلمه به بردار روش تعبیه کردن کلمات است. روش GLoVe [۱۵]، روش CBOW و skip-gram [۱۳] از بهترین نمونه‌های تعبیه کلمات هستند. ما در این تحقیق، این سه مدل را برای کار کردن با کلمات فارسی بروزرسانی کرده‌ایم. که در نهایت برای هر مدل ما دو فایل خروجی تهیه نموده ایم. فایل اول مربوط به دایره واژگان است که در آن به هر کلمه یک شناسه اختصاص داده شده و فایل دوم مربوط به بردارهای کلمات است که بردار هر کلمه با شناسه آن مشخص شده است. سپس این بردارهای کلمات به عنوان ورودی‌های شبکه عصبی RNN مورد استفاده قرار گرفته اند.

۳ مدل برچسب‌زنی اجزای سخن زبان فارسی با استفاده از شبکه عصبی بازگشتی RNN

در این تحقیق از ترکیبی از دو مدل ارائه شده در دو مقاله [۱۱] و [۱۲] استفاده شده‌است. این پیاده‌سازی با زبان پایتون و با استفاده از کتابخانه TensorFlow انجام گرفته‌است [۱۶]. در دو مقاله فوق ابتدا به بازنمایی سطح کاراکتر کلمات پرداخته شده است ولی در این تحقیق بازنمایی سطح کاراکتر مورد استفاده قرار نگرفته‌است. مدل کلی این تحقیق از سه قسمت اصلی تشکیل می‌شود:

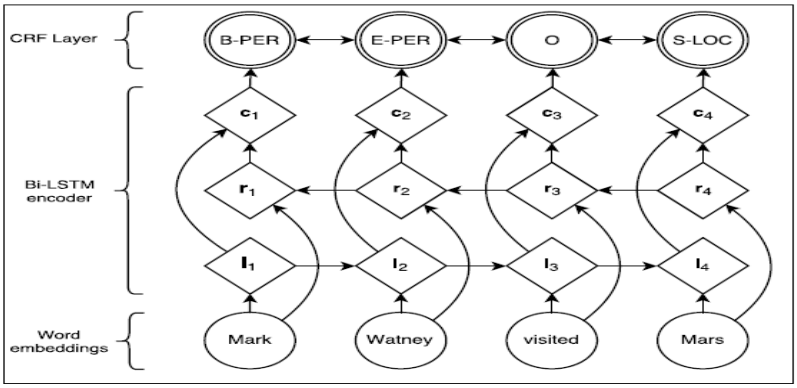
- **بازنمایی کلمات:** برای اعمال جملات به ورودی شبکه‌های عصبی نیاز است که کلمات با بردارهای عددی نشان داده‌شوند. در این تحقیق چهار نوع بردار کلمه را ارزیابی خواهیم کرد که عبارتند از بردارهای تصادفی، بردارهای CBOW، بردارهای skip-gram و بردارهای GloVe.
- **بازنمایی زمینه کلمات:** بردارهای که از مرحله قبل حاصل می‌شوند چیزی در مورد کلمات اطراف کلمه مورد نظر بیان نمی‌کنند (کلمات همسایه در جمله مورد نظر)، به همین دلیل به تولید یک بردار جدید با توجه به کلمات اطراف کلمه مورد نظر پرداخته می‌شود. تولید بردارهای هر کلمه با توجه به کلمات اطرافش باید با بردار بازنمایی معناداری نمایش داده‌شود. در این معماری برای تولید بردارها از یک شبکه LSTM استفاده می‌شود.

- **پیش‌بینی دنباله برچسب:** در مرحله آخر بردارهای کلمات و برچسب آنها در داده‌های یادگیری برای یادگیری نظارت شده مورد استفاده قرار می‌گیرند. با استفاده از شبکه آموزش داده شده، می‌توان برای هر کلمه برچسب آن را پیش بینی کرد. زمانی که به مرحله آخر می‌رسیم، برای هر کلمه یک بردار داریم که شامل اطلاعاتی از معنی کلمه و زمینه آن است. در این مقاله مجموعه برچسب‌ها شامل ۳۱ برچسب است. بنابراین ۳۱ کلاس برچسب در نظر گرفته شده‌است و برای هر کلمه یک بردار ۳۱ بعدی استفاده می‌شود. هر خانه از بردار S، نمره هر کدام از کلاس‌ها را برای آن کلمه نشان می‌دهد به عبارت دیگر خانه S_i نمره برچسب i برای کلمه مورد نظر است.

در شکل ۱ معماری اصلی شبکه عصبی مورد استفاده نمایش داده شده‌است. بردارهای تعبیه شده کلمات به یک LSTM دو طرفه داده شده‌اند که در آن l_i نمایش برداری (سطح جمله) از کلمه i و متن سمت چپ آن را ارائه می‌کند همچنین r_i نمایش برداری (سطح جمله) از کلمه i و متن سمت راست آن را ارائه می‌کند. در آخر خروجی ci از الحاق دو بازنمایی‌های چپ و راست بدست می‌آید. و بردارهای ci برای پیش‌بینی برچسب کلمه به لایه CRF داده می‌شوند.

۴. آزمایش‌ها و نتایج

در جدول ۱ مشخصات سخت افزار مورد استفاده برای اجرای برنامه‌ها را مشاهده می‌کنیم. برای آموزش شبکه عصبی ۱۰ درصد از داده را به عنوان داده آزمایش، ۲۰ درصد به عنوان validation و ۷۰ درصد را به عنوان داده آموزش استفاده کرده‌ایم. در جدول ۲ شرایط کلی آزمایش را مشاهده می‌کنیم. شبکه RNN مورد استفاده برای برچسب‌زنی با چهار نوع بردار زیر مورد ارزیابی قرار می‌گیرد:



شکل ۱. معماری اصلی شبکه عصبی بازگشتی LSTM [۱۲]

- **RNN-rand**: در این آزمایش بردارهای ورودی در ابتدا به صورت تصادفی مقداردهی می‌شوند و در طول آموزش شبکه عصبی این بردارها بروزرسانی می‌شوند. این مدل مشابه روشی است که در [۱۷] استفاده شده است.
- **RNN-skip**: در این روش بردارهای بدست آمده از مدل skip-gram را به عنوان ورودی به شبکه عصبی می‌دهیم.
- **RNN-CBOW**: در این روش بردارهای بدست آمده توسط مدل CBOW را به عنوان ورودی به شبکه عصبی می‌دهیم.
- **RNN-GLoVE**: در این روش بردارهای بدست آمده توسط مدل GLoVe را به عنوان ورودی به شبکه عصبی می‌دهیم.

پس از بکار بردن چهار نوع بردار فوق، شبکه RNN مورد استفاده از نظر دقت پیش‌بینی برچسب دستوری، مورد بررسی قرار می‌گیرد. با یک نمودار نتایج بدست آمده برای چهار نوع بردار در شکل ۲ مشاهده می‌شود. همانطور که در شکل ۲ می‌بینیم بردارهای CBOW (RNN_CBOW) روی شبکه RNN بهتر از بقیه بردارها بوده و در بیشترین تعداد چرخه آموزش (۲۰ ایپاک) دقت به ۹۷/۳۶ رسیده است.

جدول ۱. مشخصات سخت افزار

CPU	RAM	گرافیک
Intel(R)Core(TM) i3-3/70 GH	4 GB	NVIDIA- Geforce- GT720

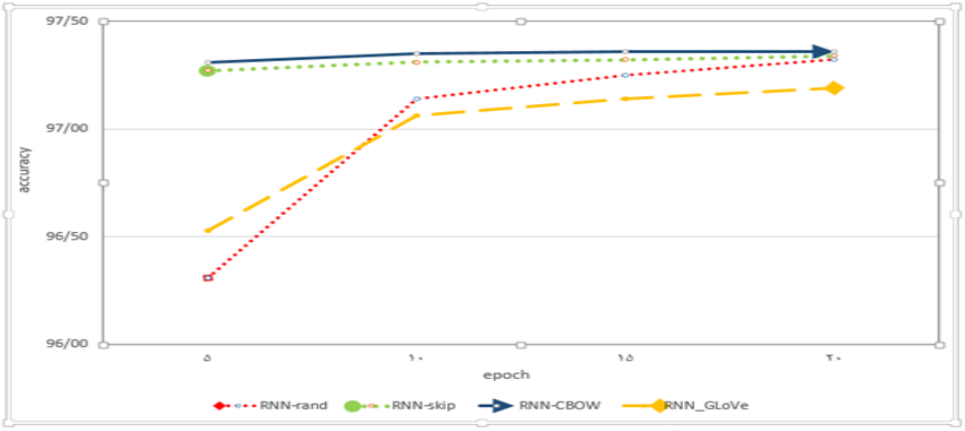
جدول ۲. شرایط کلی آزمایش

Validate	Test	Train	Batch size
260,000 words 9,119 sentences	520,000 words 16,545 sentences	1,800,000 words 66718 Sentences	32

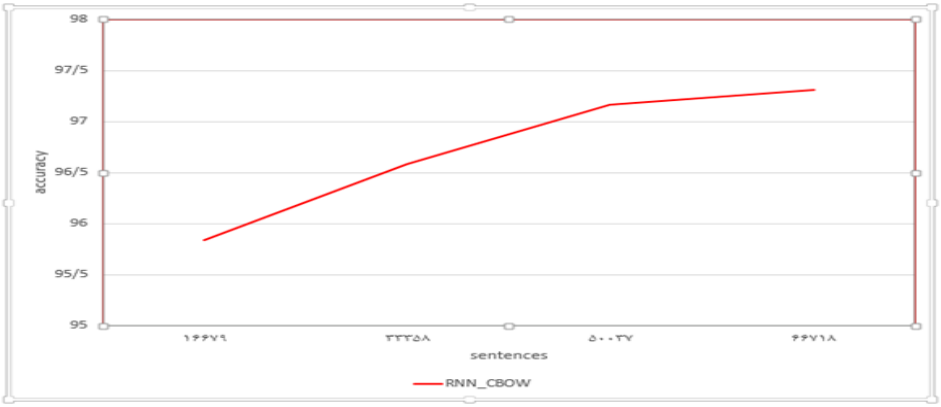
با توجه به شکل ۲ مشخص است که بردارهای CBOW بهتر از تمامی بردارها بر روی شبکه RNN عمل کرده‌اند. با در نظر گرفتن این نتیجه و با استفاده از بردارهای CBOW، در مراحل بعد به بررسی تاثیر مقدار داده آموزشی و همچنین زمان آموزش بر روی دقت این مدل می‌پردازیم. کل داده آموزشی ۶۶۷۱۸ جمله دارد. در این آزمایش کل داده آموزشی به چهار قسمت تقسیم شده‌است و نمودار تاثیر مقدار داده آموزشی (تعداد جملات) بر دقت مدل را در شکل ۳ می‌بینیم. در این آزمایش تعداد دسته‌های آموزشی برابر با ۳۲ و تعداد چرخه‌های آموزش (epoch) برابر با ۵ است.

در شکل ۳ تغییرات دقت نسبت به تعداد جملات آموزشی مشاهده می‌شود. دیده می‌شود که با ۱۶۶۷۹ جمله به دقت ۹۵/۸۴ دست یافته‌ایم و در بهترین حالت با تعداد جملات ۶۶۷۱۸ دقت به ۹۷/۳۱ رسیده است. بنابراین تاثیر تعداد جملات داده در دقت شبکه قابل توجه ارزیابی می‌شود و احتمالا با افزایش داده‌های آموزشی دقت شبکه باز هم بهتر شود.

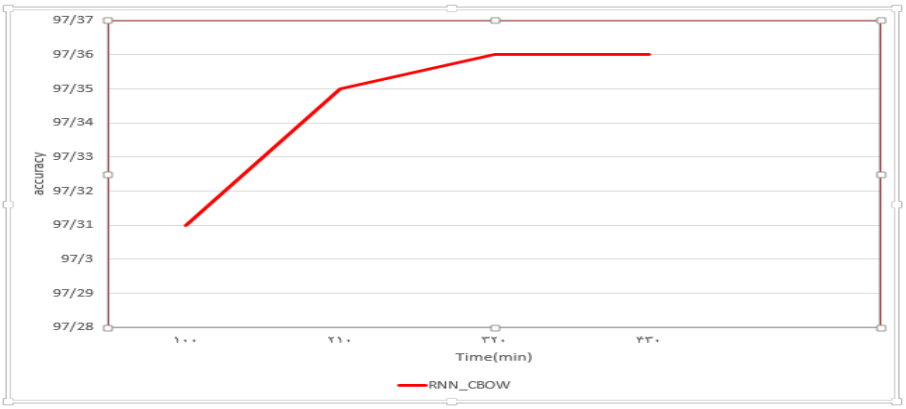
در این مدل برای اجرای هر ایپاک تقریبا ۲۱ دقیقه زمان نیاز است (به عنوان مثال ۲۰ ایپاک به طور تقریبی در ۴۳۰ دقیقه اجرا می‌شود). در شکل ۴، ارتباط زمان آموزش مدل RNN با دقت شبکه در بردارهای CBOW مورد بررسی قرار گرفته است که در این نمودار محور افقی مربوط به زمان اجرا به ترتیب در ۵، ۱۰، ۱۵ و ۲۰ ایپاک است. در این نمودار زمان با دقیقه نشان داده شده است. مشاهده می‌شود که شبکه در زمان ۱۰۰ دقیقه به دقت ۹۷/۳۱ و در زمان ۴۳۰ دقیقه به دقت ۹۷/۳۶ رسیده است. با توجه به مسطح شدن تقریبی انتهای نمودار می‌توان نتیجه گرفت که مدل به اندازه کافی آموزش دیده است و صرف زمان بیشتر بهبود قابل توجهی در نتایج ایجاد نخواهد کرد.



شکل ۲. نمودار نتایج بردارهای مختلف در شبکه عصبی



شکل ۳. نمودار دقت بردارهای CBOW نسبت به اندازه داده‌ها



شکل ۴. نمودار دقت بردارهای CBOW نسبت به زمان

منابع

[1]Seraji, M., B. Megyesi, and J. Nivre. *A basic language resource kit for Persian*. in *Eight International Conference on Language Resources and Evaluation (LREC 2012), 23-25 May 2012, Istanbul, Turkey*. 2012. European Language Resources Association.

[۲]Seraji, M. *A statistical part-of-speech tagger for Persian*. in *NODALIDA 2011, Riga, Latvia, May 11–13, 2011*. 2011.

[۳]Bengio, Y., P. Simard, and P. Frasconi, *Learning long-term dependencies with gradient descent is difficult.* IEEE transactions on neural networks, 1994. **5**(2): p. 157-166.

[۴]Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.

[۵]Cho, K., et al., *Learning phrase representations using RNN encoder-decoder for statistical machine translation.* arXiv preprint arXiv:1406.1078, 2014.

[۶]Kalchbrenner, N. and P. Blunsom. *Recurrent Continuous Translation*

Models. in *EMNLP*. 2013.

[۷] Graves, A., *Generating sequences with recurrent neural networks.* arXiv preprint arXiv:1308.0850, 2013.

[۸] Bahdanau, D., K .Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate.* arXiv preprint arXiv:1409.0473, 2014.

[۹] Sutskever, I., O. Vinyals, and Q.V. Le. *Sequence to sequence learning with neural networks*. in *Advances in neural information processing systems*. 2014.

[۱۰] Huang, Z., W. Xu, and K. Yu, *Bidirectional LSTM-CRF models for sequence tagging.* arXiv preprint arXiv:1508.01991, 2015.

[۱۱] Ma, X. and E. Hovy, *End-to-end sequence labeling via bi-directional lstm-cnns-crf*. arXiv preprint arXiv:1603.01354, 2016.

[۱۲] Lample, G., et al., *Neural architectures for named entity recognition.* arXiv preprint arXiv:1603.01360, 2016.

[۱۳] Mikolov, T., et al., *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.۲۰۱۳ ,

[۱۴] Bijankhan, M., *The role of the corpus in writing a grammar: An introduction to a software.* Iranian Journal of Linguistics, 2004. **19**.(۲)

[۱۵] Pennington, J., R. Socher, and C. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

[۱۶] *Named Entity Recognition (LSTM + CRF) - Tensorflow* n.d. [cited 2017 10/9]; NER in Tensorflow with LSTM and CRF]. Available from: https://github.com/guillaumegenthial/sequence_tagging.

[۱۷]Collobert, R., et al., *Natural language processing (almost) from scratch.* Journal of Machine Learning Research, 2011. 12(Aug): p. 2493-2537.

رایسست‌کیوترنسلیت: یک نظام ماشین ترجمه مبتنی بر پیشنهاد جهت بازیابی اطلاعات بین زبانی انگلیسی و فارسی در زمینه پزشکی

امین رحمانی*، محمدرضا فلاحتی قدیمی فومنی و محمدباقر دستغیب*****

چکیده

در سال‌های اخیر، تولید اطلاعات در فضای مجازی با سرعت چشمگیری در حال افزایش می‌باشد. تولید اطلاعات فقط به زبان انگلیسی محدود نمی‌گردد، بلکه در زبان‌های دیگر نیز تولید محتوا صورت می‌گیرد. جهت دسترسی به این محتوای چندزبانه، باید از ابزارهای هوشمندی مانند نظام‌های بازیابی اطلاعات بین زبانی استفاده نمود. بر این اساس، در پژوهش حاضر، یک رویکرد جدید برای بازیابی اطلاعات بین زبانی انگلیسی و فارسی بررسی شده است. در دهه‌های اخیر، پژوهش‌هایی در زمینه بازیابی اطلاعات بین زبانی انجام شده است، ولی هنوز مسائل حل نشده زیادی در این زمینه وجود دارد. هدف از انجام پژوهش حاضر ساخت یک نظام بازیابی اطلاعات بین زبانی به نام رایسست‌کیوترنسلیت[1] می‌باشد. این نظام هوشمند کاربران غیر فارسی زبان را قادر می‌سازد تا از منابع فارسی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری نیز استفاده نمایند. در این پژوهش دو الگوریتم جدید ارائه شده است. الگوریتم اول، تحت عنوان EPATA1[2] یک الگوریتم ترجمه متن خودکار می‌باشد. الگوریتم دوم، یک الگوریتم تغییر اولویت پیشنهاد پرس‌وجو می‌باشد. پس از ارزیابی نظام تولید شده، الگوریتم ترجمه به صورت معناداری نمره بلو را افزایش داد. همچنین نمره آزمون میانگین متوسط دقت نیز افزایش یافت.

واژه‌های کلیدی: بازابی اطلاعات بین زبانی، رایسست‌کیوترنسلیت، EPATA1، نمره بلو، میانگین متوسط دقت

---

* دانشجوی کارشناسی ارشد زبانشناسی رایانشی، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، A.Rahmani@ricest.ac.ir

** استادیار گروه پژوهشی زبانشناسی رایانشی، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری (نویسنده مسئول)، mrfalahat@yahoo.com

*** استادیار گروه پژوهشی طراحی و عملیات سیستمها، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، Hdastgheib@gmail.com

[1] رایسست‌کیوترنسلیت نظام بازیابی اطلاعات بین زبانی طراحی شده در پژوهش حاضر می باشد.

[2] English-Persian Automatic Translation Algorithm1

۱. مقدمه

فضای اینترنت، یک حافظه عظیم و چندزبانه از اطلاعات به شمار می‌رود. از زمان پیدایش، اطلاعات در این فضای مجازی با سرعت بسیاری رشد پیدا کرده است. اطلاعات در این فضا، به زبان‌های مختلفی تولید شده است. نیاز به بازیابی اطلاعات چندزبانه در فضای مجازی انگیزه‌ای برای به وجود آمدن نظام‌های بازیابی اطلاعات بین زبانی به شمار می‌رود. اگر یک پرس‌وجو و اسناد بازیابی شده به یک زبان باشند، این فرایند تحت عنوان بازیابی اطلاعات تک زبانه مطرح می‌شود، اما اگر پرس‌وجوی وارد شده به یک زبان باشد و اسناد بازیابی شده به زبان دیگر این فرایند تحت عنوان بازیابی اطلاعات بین زبانی شناخته می‌شود[1,2]. به علت وجود موانع زبانی، دسترسی به اطلاعات چندزبانه مشکل می‌باشد و نظام‌های بازیابی اطلاعات بین زبانی این امکان را فراهم می‌سازد تا پرس‌وجو را، مثلا، به یک زبان وارد، و سپس اطلاعات به زبان دیگری بازیابی شود[3,4]. از زمان تاسیس، در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری (رایسست)[1]، ده‌ها پایگاه اطلاعاتی فارسی و انگلیسی تولید شده است. به علت وجود وجود موانع زبانی، کاربران غیر فارسی زبان قادر به استفاده از پایگاه‌های فارسی این مرکز بین‌المللی نیستند. برای حل این مشکل، در پژوهش حاضر یک نظام بازیابی اطلاعات بین زبانی تحت عنوان رایسست‌کیوترنسلیت طراحی گردید تابه کاربران غیر فارسی زبان کمک نماید تا از پایگاه‌های اطلاعاتی فارسی رایسست استفاده نمایند.

رویکردهای مختلفی نسبت به طراحی نظام‌های بازیابی اطلاعات بین زبانی وجود دارد. از جمله این رویکردها می‌توان به «ترجمه پرس‌وجو»، «ترجمه سند» و «ترجمه پرس‌وجو» و سند اشاره کرد[5,6]. هر کدام از این مدل‌ها معایب و محاسن مختلفی دارند. مثلا روش‌های ترجمه سند و ترجمه پرس‌وجو و سند هزینه‌ی محاسباتی زیادی را به سیستم وارد می‌کنند. به همین دلیل در سال‌های اخیر ترجمه پرس‌وجو بیشتر مورد استفاده قرار گرفته است[7]. واژه‌نامه ماشین‌خوان، پیکره دوزبانه و ماشین ترجمه ابزارهایی هستند که در ترجمه پرس‌وجو می‌توان از آن‌ها بهره گرفت. واژه‌نامه و پیکره دوزبانه مشکل عدم پوشش‌دهی کامل کلمات را دارند؛ به همین دلیل از ماشین ترجمه در ترجمه پرس‌وجو بیشتر استفاده شده است[8,9]. اگرچه ماشین ترجمه‌های آماری از انواع دیگر ماشین ترجمه مانند مبتنی‌برقانون و

1 WWW.RICeST.ac.ir

مبتنی‌برمثال از دقت بالاتری برخوردار هستند، اما الگوریتم ترجمه آماری هزینه‌ی محاسباتی بسیار بالایی دارد[10]. ادوپا و ماجی [11] براین باورند که جهت آموزش الگوریتم آماری داده بسیار حجیمی نیاز است. علاوه بر داده حجیم، این الگوریتم بسیار پیچیده بوده و از نظر هزینه محاسبه الگوریتمی، هزینه الگوریتم چندجمله‌ای را در بردارد. در پژوهش حاضر یک الگوریتم ترجمه خودکار تحت عنوان EPATA1 پیشنهاد شده است که از لحاظ پیچیدگی زمانی بسیار از الگوریتم‌های آماری ساده‌تر است و در عین سادگی از دقت لازم در ترجمه نیز برخوردار می‌باشد. در بخش‌های بعدی در مورد این الگوریتم بحث خواهد شد. در بخش بعد پژوهش‌های مرتبط مورد بررسی قرار می‌گیرد.

۱-۱.پژوهش‌های انجام شده در زمینه بازیابی اطلاعات بین زبانی

از سال ۱۹۹۶ اولین تلاش‌ها در زمینه طراحی و ساخت نظام‌های بازیابی اطلاعات بین زبانی صورت گرفت[12]. ابزارهای مختلفی مانند واژه‌نامه، مدل‌های احتمالی و ماشین ترجمه در این فرایند به کار گرفته شده است. صالح و پسینا [2] با استفاده از یک روش یادگیری ماشین جدید و یک ماشین ترجمه آماری یک نظام بازیابی اطلاعات بین زبانی طراحی نمودند. شارما و مارول [5] نیز به بررسی مدل‌ها و پژوهش‌های مرتبط در این حوزه پرداختند. نیکولینا و همکاران [13] از ماشین ترجمه جهت ترجمه پرس‌وجو استفاده نمودند. علاوه مدل آماری، این پژوهشگران با اضافه کردن مشخصه‌های دستوری به فرایند ترجمه، الگوریتم پیشنهادی را بهینه نمودند. ترو و همکاران [9] نیز روش‌های مختلف ترجمه آماری را در ترجمه پرس‌وجو باهم ترکیب کردند. نتیجه پژوهش رضایت‌بخش گزارش شد. گوپتا و همکاران [14] نیز با استفاده از یک ماشین ترجمه آماری روال بازیابی اطلاعات بین زبانی را تسریع نمودند. نتیجه پژوهش نشان داد که استفاده از ماشین ترجمه در مقایسه با دیگر ابزارها، نتیجه بازیابی اطلاعات را بهبود می‌بخشد. نوگویان و همکاران [15] از ویکیپدیا به عنوان داده آموزش، استفاده نمودند. پژوهشگران با استفاده از یک ماشین ترجمه آماری عملیات ترجمه پرس‌وجو را انجام دادند. عملکرد این نظام ۶۷ درصد شبیه به نظام بازیابی تک زبانه گزارش شد. جدول ۱ پژوهش‌های دیگر را نشان می‌دهد:

جدول ۱. خلاصه‌ای از پژوهش‌های انجام شده

سال	نویسنده (گان)	ابزار ترجمه	نتایج
2017	P. Iswarya And V. Radha [3]	hybrid machine translation	Optimization over MAP score.
2017	Rakshita Rao and Mangala Madankar[1]	Machine translation	Improvements over preexisting monolingual system.
2017	Rahmani, Falahati, Dastgheib [16]	Google Translate	Improvement over preexisting monolingual information retrieval system.
2016	Paheli Bhattacharya, Pawan Goyal and Sudeshna Sarkar[4]	Multilingual Word Clusters	Better results than preexisting model.
2013	Saravanan et al. [17]	Bilingual dictionary +Enhanced Transliteration	Hindi to Eng MAP score:0.4977 & Tamil to Eng MAP score:0.4145
2011	Manikandanand Shriram [18]	Bilingual dictionary	An effective strategy to implement query translation
2010	Pemawat et al. [19]	Bilingual dictionary	Improvements over precision and recall scores.
2010	Antony et al. [20]	Parallel corpus	Proposed model gives better results than existing.
2010	Saraswathi et al. [21]	Machine Translation	Tamil MAP score Increased by 60%.

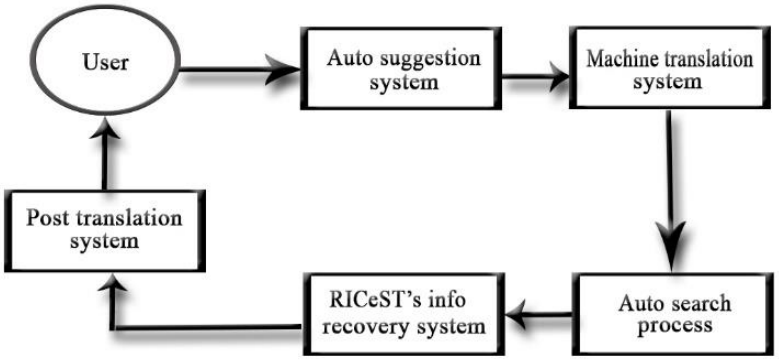
جدول ۱ شامل پژوهش‌های انجام شده در زمینه بازیابی اطلاعات بین زبانی می‌باشد. در ستون اول سال و در ستونهای بعدی به ترتیب نام نویسندگان، ابزار ترجمه پرس‌وجو و نتایج گنجانده شده است. در ادامه به بررسی روش‌شناسی پژوهش حاضر پرداخته خواهد شد.

۲. روش تحقیق

در این بخش به بررسی رویکرد پیشنهادی در این پژوهش، داده مورد استفاده، معماری نظام رایسست‌کیوترنسلیت، الگوریتم ترجمه اتوماتیک و بخش‌های دیگر این پژوهش پرداخته خواهد شد.

۲-۱. رویکرد پیشنهادی بازیابی اطلاعات بین زبانی

در این بخش به بررسی رویکرد پیشنهادی در پژوهش حاضر پرداخته خواهد شد. در تصویر ۱ رویکرد مورد نظر نشان داده شده است:



تصویر۱. رویکرد بازیابی اطلاعات بین زبانی

به محض ورود پرس‌وجو توسط کاربر در بخش ترجمه و جستجوی پرس‌وجوی نظام رایسست‌کیوترنسلیت، پنج پرس‌وجویی که بیشترین ارتباط را با نویسه‌های وارد شده دارد به کاربر پیشنهاد داده می‌شود. سپس کاربر یا پرس‌وجوی پیشنهاد داده شده را انتخاب می‌نماید و یا یک پرس‌وجو را خود تولید نموده و وارد می‌نماید. پس از ورود، پرس‌وجو توسط نظام ترجمه خودکار از انگلیسی به فارسی ترجمه می‌گردد. سپس پرس‌وجوی ترجمه شده در پایگاه داده فارسی رایسست جستجو شده و در نهایت تعدادی سند مرتبط با پرس‌وجو بازیابی می‌گردد. پس از عملیات بازیابی اطلاعات، نظام ترجمه متن به کاربر این امکان را می‌دهد تا عناوین اسناد بازیابی شده را از فارسی به انگلیسی ترجمه نماید.

۲-۲. داده پژوهش

بنابر گزارش اخذ شده از گروه پژوهشی طراحی و عملیات سیستم‌ها، در تاریخ ۱۰ خرداد ۱۳۹۶ حدود ۷۵۰۰۰ مقاله در حوزهٔ پزشکی در پایگاه اطاعاتی رایسست موجود بوده است. به دلیل آزمایشی بودن این پژوهش، ۲۵۰ عنوان انگلیسی مقاله به عنوان داده نظام پیشنهاد پرس‌وجو به روش تصادفی ساده انتخاب شدند. از این ۲۵۰ عنوان، حدودا ۱۲۰۰ کلیدواژه و عبارت معنادار به صورت دستی استخراج گردید که از این داده استخراج شده در نظام پیشنهاد پرس‌وجو استفاده شد. حدود ۱۵۰۰۰ عنوان مقاله فارسی و انگلیسی به همراه چکیده، به عنوان داده ماشین ترجمه در این پژوهش مورد استفاده قرار گرفت.

طبق مشاهدات و اطلاعات استخراج شده از مقالات مرتبط، داده مرسوم جهت ارزیابی نظام‌های بازیابی اطلاعات بین زبانی «مجموعه داده ترک[1]» می‌باشد. این داده شامل پرس‌وجوهایی از پیش تعیین شده در حوزه‌های مختلف علمی است. در این پژوهش، از داده پزشکی ترک که در سال ۲۰۱۲ توسط سایت ترک[2] انتشار یافته، استفاده شد. پژوهشگر از کلمات و جملات این مجموعه داده جهت ارزیابی نظام بازیابی اطلاعات استفاده نمود. در مجموع، ۵۰ کلمه و عبارت دو تا پنج کلمه‌ای استخراج شده از داده ترک پرس‌وجوهای مورد استفاده این پژوهش را تشکیل داد. در جدول ۲ نمونه پرس‌وجوهای استخراج شده از داده ترک نشان داده شده است:

جدول ۲. نمونه پرس‌وجوهای استخراج شده از داده ترک

ردیف	پرس‌وجوهای انگلیسی	معادل فارسی
۱	Children with dental caries	کودکان مبتلابه پوسیدگی دندانی
۲	inflammatory disorders	اختلالات التهابی
۳	tubular necrosis	نکروز لوله‌ای
۴	blood coagulation	انعقاد خون
۵	intravascular	داخل عضلانی
۶	Alzheimer disease	بیماری آلزایمر

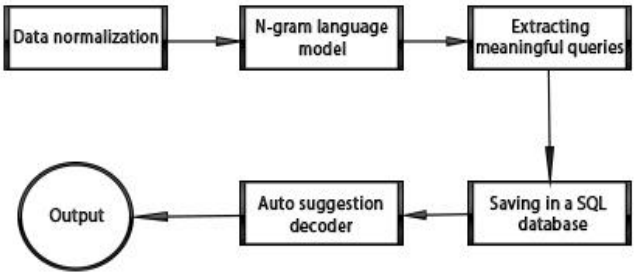
1 TREC
2 http://trec.nist.gov/data/medical/12/topics136-185.txt

ردیف	پرس‌وجوهای انگلیسی	معادل فارسی
۷	endarterectomy	آندروترکتومی
۸	diabetes mellitus	دیابت قندی
۹	thrombocytosis	ترومبوسیتوز
۱۰	lupus nephritis	سل جلدی نفریت

این پرس‌وجوها استخراج شده توسط یک پزشک ترجمه گردید. همچنین، پرس‌وجوهای انگلیسی و معادل فارسی ترجمه شده توسط دو متخصص به صورت کامل کنترل شد.

۲-۳. معماری نظام رایسست‌کیوترنسلیت

در این بخش معماری و نحوه ساخت نظام بازیابی اطلاعات بین زبانی رایسست‌کیوترنسلیت بررسی می‌شود. تصویر ۲ نشان دهنده بخش‌های مختلف نظام رایسست‌کیوترنسلیت می‌باشد:



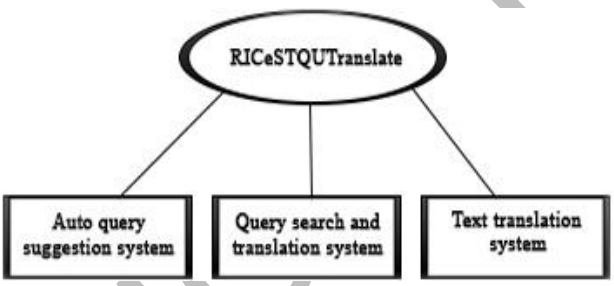
تصویر۲. معماری رایسست کیوترنسلیت

رایسست‌کیوترنسلیت دارای سه بخش اصلی می‌باشد. بخش اول تحت عنوان نظام پیشنهاد پرس‌وجو وظیفه تسهیل عملیات ورود پرس‌وجو توسط کاربر را به عهده دارد. این پرس‌وجوها از قبل استخراج و در حافظه این بخش ذخیره شده است. بخش دوم تحت عنوان نظام ترجمه و جستجوی پرس‌وجو، عبارت‌های ورودی توسط کاربر را از انگلیسی به فارسی ترجمه کرده و در نظام رایسست سرچ می‌نماید. بخش سوم نیز تحت عنوان نظام ترجمه متن، وظیفه عناوین اسناد بازیابی شده را از فارسی به انگلیسی برعهده دارد. در بخش‌های بعدی

نحوه ساخت قسمت‌های مختلف رایسست‌کیوترنسلیت بررسی می‌شود.

۲-۳-۱. بخش پیشنهاد پرس‌وجو

پیشنهاد پرس‌وجو روشی است جهت کمک به کابران تا بدینوسیله هم اسناد مرتبط با پرس‌وجو را بازیابی کنند و هم ازهدر رفتن وقت جلوگیری شود. کیم و همکاران [22] معتقدند که این روش تاثیر بسزایی در دقت در بازیابی اطلاعات دارد. همچنین، با توجه به آماده بودن پرس‌وجوها بسیار به کاربر کمک می‌کند. " به علت این که در این روش پرس‌وجوها آماده هستند روش پیشنهاد پرس‌وجو یک متد موثر در کمک به کاربران به حساب می‌آید"[23,24].
در تصویر ۳ مدل پیشنهادی جهت طراحی نظام پیشنهاد پرس‌وجو نشان داده شده است:



تصویر۳. مدل پیشنهادی جهت طراحی نظام پیشنهاد پرس‌وجو

در مرحله اولیه ساخت نظام پیشنهاد پرس‌وجو، داده پژوهش از نظر درستی کلمات و نکات نگارشی کنترل شد. به‌دلیل وجود غلط‌های املایی و نگارشی احتمالی تمام ۲۵۰ عنوان مقاله مورد بررسی قرار گرفت. غلط‌های نگارشی عناوین انگلیسی به صورت خودکار توسط نرم‌افزار مایکروسافت ورد ۲۰۱۳[1] کنترل و رفع شد.

پس از نرمال‌سازی عناوین، 5-gram تمام جملات انگلیسی توسط آخرین نسخه نرم‌افزار آنت‌کونک[2] به صورت اتوماتیک استخراج شد. نسخه 4.3.3 این نرم‌افزار در سال ۲۰۱۴ طراحی شد. این نرم‌افزار قادر است به‌غیراز n-gram، کنکوردنس کلمات[3] را نیز استخراج بنماید. تصویر

1 MS Word 2013
2AntConc
3 concordance

۳-۱ نشان دهندهٔ شمای کلی این نرم‌افزار است. ابزار تحلیل متن مانند n-gram و دیگر ابزارها به راحتی در این ابزار قابل استفاده می‌باشد. علاوه بر ابزارهای مذکور، این نرم‌افزار قادر است باهم آیی کلمات[1] را نیز استخراج نماید که در تحلیل متون بسیار کاربرد دارد. نکته قابل ذکر برای استفاده از این ابزار این است که فایل‌های ورودی به این نرم‌فزار باید با روش کدگذاری یوتی‌اف هشت[2] ذخیره شده باشند، در غیر این صورت کاراکترهای فارسی برای این نرم‌فزار قابل قابل شناسایی نیست. پس‌از انجام عملیات 5-gram، تمام کلمات انگلیسی استخراج شده، تمامی کلمات و عبارات معنادار به صورت کاملاً دستی و توسط پژوهشگر از این جدول استخراج شد و سپس در یک جدول در پایگاه‌داده اس‌کیوال سرور[3]، بر اساس بسامد کلمات و عبارات ذخیره گردید. این نکته قابل ذکر است که، این جدول حافظه نظام پیشنهاد پرس‌وجو محسوب می‌گردد. در ادامه رمزگشای نظام پیشنهاد پرس‌وجو بررسی می‌گردد.

همانطور که بیان شد پرس‌وجوهای استخراج شده در یک پایگاه داده اس‌کیوال سرور ذخیره شده‌است. یک رابط‌گرافیکی وظیفه برقراری ارتباط میان کاربر و حافظه نظام پرس‌وجو را برعهده دارد. این رابط گرافیکی همان رمزگشای نظام پرس‌وجو است. کاربر با وارد کردن پرس‌وجو در یک جعبه متنی، تعدادی گزینه پیش‌رو دارد، که قادر است از میان این پرس‌وجوها یکی را انتخاب نماید. در این پژوهش فقط از پرس‌وجوهای انگلیسی استفاده شده است که البته در صورت نیاز امکان ارتقاء نظام مذکور وجود دارد و می‌توان نظام پیشنهاد پرس‌وجوی فارسی را نیز به این ابزار اضافه نمود. همانطور که قبلاً نیز مطرح شد، این پژوهش در مقیاس آزمایشی انجام شده است و از این رو به ایجاد نظام پیشنهاد پرس‌وجو به زبان انگلیسی بسنده شده است. این رابط گرافیکی بوسیله زبان سی‌شارپ طراحی گردیده است. در تصویر ۴ نمایی از پیشنهاد پرس‌وجو قابل مشاهده است :

1 Collocation
2 UTF-8
3 SQL server

تصویر۴. نمایی از پیشنهاد پرس‌وجو

به محض ورود پرس‌وجو توسط کاربر، تعدادی گزینه به کاربر نمایش داده می‌شود، و نرم‌افزار امکان انتخاب یک گزینه را به کاربر می‌دهد. هرچه بسامد اولیه پرس‌وجوی ذخیره شده در حافظه نظام پیشنهاد پرس‌وجو بیستر باشد در جایگاه بالاتری به کاربر پیشنهاد داده می‌شود. اما روند اولویت پیشنهاد پرس‌وجو به صورت یکنواخت صورت نمی‌گیرد و بعد از استفاده از نظام رایسست‌کیوترنسلیت در طول زمان اولویت پیشنهاد پرس‌وجوها تغییر می‌کند. الگوریتم تغییر اولویت پیشنهاد در پژوهش حاضر براساس دو معیار فرکانس اولیه و ضریب انتخاب عمل می‌کند. فرمول ۱ نشان‌دهنده‌ی نحوه تغییر اولویت پیشنهاد پرس‌وجو می‌باشد:

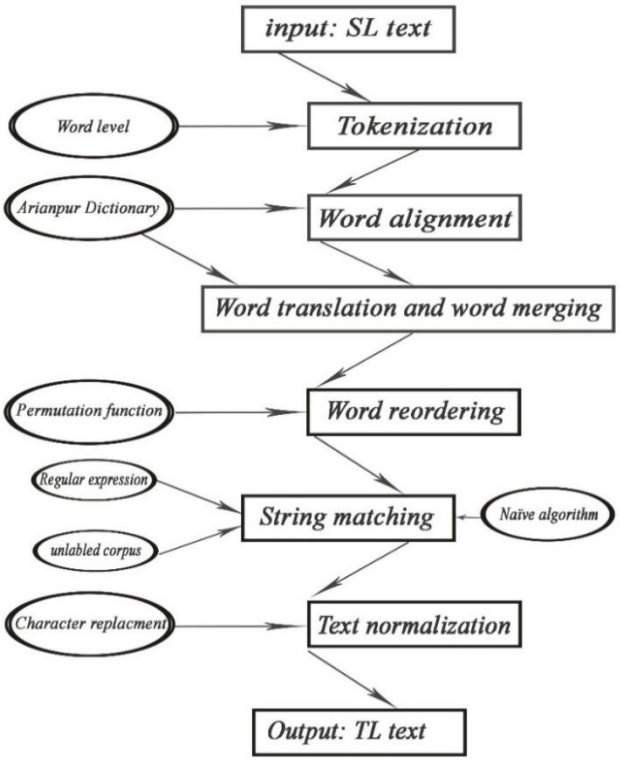
$$\text{Suggestion priority} = \frac{\text{SC+IF}}{MAX\,(SC)+MAX\,(IF)} \qquad (۱)$$

در فرمول ۱، SC همان ضریب انتخاب پرس‌وجو توسط کاربر است و IF همان بسامد اولیه مربوط به هر پرس‌وجو می‌باشد. با انتخاب هر پرس‌وجو مقدار SC تغییر کرده و باعث می‌شود که اولویت پیشنهاد پرس‌وجوی مورد نظر نیز تغییر کند. رابطه فوق با تقسیم کردن بیشینه[1] بدست آمده از هر دو فاکتور IF و SC نرمال شده و بین ۰ تا ۱ قرار می‌گیرد. در این پژوهش با آزمایش بر روی داده واقعی این رابطه به دست آمده است.

1 MAX

۲-۳-۲. بخش ترجمه و جستجوی پرس‌وجو

در این بخش به بررسی الگوریتم ترجمه خودکار پیشنهاد شده و همچنین نظام ترجمه و جستجوی پرس‌وجو پرداخته خواهد شد. تصویر ۵ الگوریتم ترجمه خودکار را نشان می‌دهد:



تصویر ۵. الگوریتم ترجمه خودکارEPATA1

در این قسمت اساس کار و الگوریتم ماشن ترجمه مرحله به مرحله تشریح می‌شود. در مرحله اول با ورود اطلاعات توسط کاربر این ماشین ترجمه فعال می‌گردد. با ورود متن، اطلاعات از کاربر اخذ می‌شود.

مرحلهٔ اول : تقطیع متن[1]: یکی از اولین و اساسی‌ترین کارهایی که در پردازش زبان طبیعی انجام می‌شود تقطیع یک متن یا یک صوت به کلمات یا حروف تشکیل دهنده آن است. تقطیع

1 Tokenization

در سه سطح انجام می‌پذیرد. یک متن را نسبت به طول و هدف مورد نظر می‌توان به پاراگراف، جمله و کلمه تقطیع کرد. در این پژوهش بدلیل کوتاه بودن متن ورودی توسط کاربر، این متن به کلمات تشکیل دهنده می‌شکند. متن وارد شده، به نویسه‌های زیر حساس بوده و تقطیع می‌شود.

جدول ۳. نویسه‌های تابع تقطیع متن

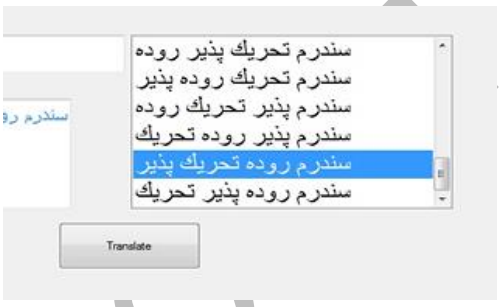
{	}	(	)
[	]	>	<
-	_	=	+
/	?	~	!
' '	&	*	.
\\	:	;	"
|	\t	\n	\r
^	,	%	$
#	@	"	،
۰	۱	۲	۳
۴	۵	۶	۸
۸	۹		

مرحلهٔ دوم: تراز کردن[1]*، ترجمه*[2] *و ادغام*[3] *کلمات*: پس از تقطیع کلمات، در مرحله اول با یافتن کلمات در پایگاه داده فرهنگ‌لغت الکترونیکی آریان‌پور[4]، کلمات مورد نظر باهم تراز می‌گردد. پس از هم‌ترازی کلمات، معادل فارسی کلمات انگلیسی در فهرستی نگهداری می‌شود تا در مرحله بعدی مورد استفاده قرار بگیرد. در نهایت، کلمات فارسی مورد نظر باهم ادغام شده و یک عبارت را تشکیل می‌دهد. نکته مهم این است که در مرحله ترجمه، از نخستین معادل پیشنهادی واژه‌نامه استفاده می‌شود.

1 Word alignment
2 Word translation
3 Word merging
4 https://aryanpour.com/

تصویر۶. تقطیع متن

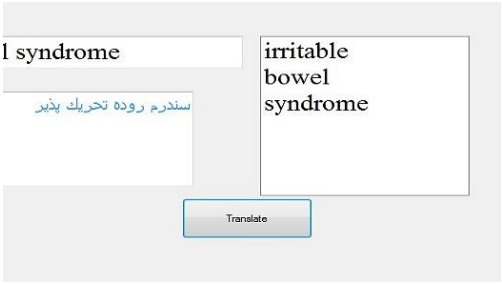
در ادامه ترجمه و ادغام کلمات با کمک تصویر نشان داده خواهد شد.



تصویر ۷. ترجمه و ادغام کلمات

مرحلهٔ سوم: جایگشت[1] *کلمات*: تا این مرحله، کلمات ترجمه شده با ترتیب کلمات زبان مبدأ یعنی انگلیسی درکنار هم قرار گرفتند، اما در این مرحله نیاز است که به نحوی ترتیب کلمات با ترتیب کلمات در زبان فارسی منطبق شوند. برای این منظور از یک تابع جایگشت یک سویه استفاده می‌شود تا ترکیبات ممکن از کلمات را تولید کند.

1 Permutation



تصویر ۸. جایگشت کلمات یک عبارت

مرحلهٔ چهارم: تطابق[1] *عبارت تولید شده*: در مرحله نهایی، عبارت تولید شده، در یک پیکره فارسی جستجو می‌شود. این جستجو توسط ماشین خودکار متناهی[2] انجام می‌گردد. پیکره موجود بدون برچسب بوده و دارای ۱۵ هزار عنوان و چکیده فارسی در زمینه پزشکی می‌باشد. پس از جستجو، عبارت مورد نظر با ترتیب کلمات زبان فارسی در متغیری ذخیره می‌گردد تا پس از تصحیح به کاربر نمایش داده شود.

مرحلهٔ پنجم: تصحیح متن[3]: در این مرحله تمامی متن تولید شده و جستجو شده، از لحاظ غلط‌های احتمالی تصحیح می‌گردد. عمدهٔ اشتباهات نگارشی در متون فارسی را می‌توان به وجود نویسه‌های عربی و رعایت نکردن نیم‌فاصله بین پیشوند و کلمه بعد از آن و پسوند و کلمه قبل از آن نسبت داد. تصویر ۹ نظام ترجمه و جستجوی پرس‌وجو را نشان می‌دهد:

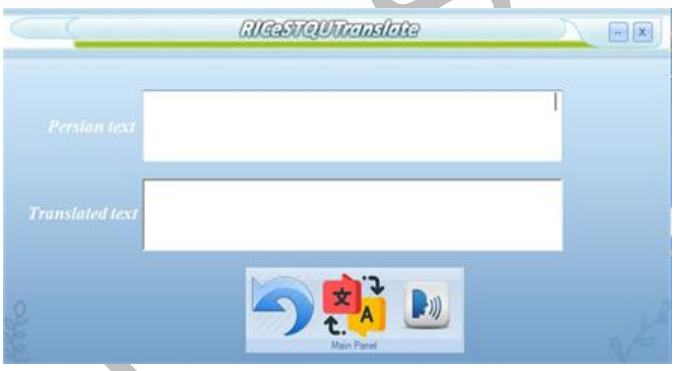
تصویر۹. محیط بخش ترجمه و جستجوی پرس‌وجو

---

1 Matching
2 Finite State Automata
3 Text normalization

۲-۳-۳. بخش ترجمه متن

هدف از طراحی این بخش، این است که کاربران بالقوه نظام رایسست‌کیوترنسلیت پس از بازیابی اطلاعات به زبان فارسی، بتوانند عناوین اسناد بازیابی شده را ترجمه کنند. این امر، به کاربران امکان می‌دهد تا اسناد بازیابی شدهٔ مرتبط با پرس‌وجوی مورد نظر را شناسایی و در نهایت متن کامل سند را از پایگاه مقالات فارسی رایسست بازیابی نمایند. اگر این ابزار در اختیار کاربران نباشد، پس از جستجوی پرس‌وجو فقط تعدادی سند بازیابی می‌شود؛ که کاربر نا آشنا به زبان فارسی در درک این اسناد با مشکل روبرو خواهد شد. جهت طراحی این ابزار از ماشین ترجمه گوگل‌ترنسلیت استفاده شده است. با وارد کردن این مترجم ماشینی به بسته برنامه‌نویسی ویژوال استودیو[1] امکان ترجمه جملات و عناوین مربوط به اسناد بازیابی شده فراهم شده است. تصویر ۱۰ شمای کلی این ماشین ترجمه را نشان می‌دهد:



تصویر ۱۰. نمای کلی بخش ترجمه متن

کاربر با وارد کردن متن در قسمت ورودی متن[2] و زدن دکمه ترجمه[3]، متن ترجمه شده[4] را در جعبه متنی متن ترجمه شده مشاهده می‌گردد.

1 Visual studio
2 Input text
3 Translate
4 Translated text

۳. نتایج

در این بخش تحلیل کیفی نرم‌افزار رایسست‌کیوترنسلیت با دو روش معمول ارزیابی نظام‌های پردازش زبان طبیعی یعنی جعبه‌شیشه‌ای و جعبه‌سیاه بررسی می‌گردد . در ابتدا نظام رایسست‌کیوترنسلیت به صورت یکپارچه ارزیابی می‌شود. سپس، در بخش تحلیل جعبه‌شیشه‌ای، ماشین ترجمه به عنوان جزء اصلی نرم‌افزار رایسست‌کیوترنسلیت ارزیابی خواهد شد.

۳-۱. ارزیابی جعبه‌سیاه

در این بخش به ارزیابی کلی نظام بازیابی اطلاعات بین زبانی رایسست‌کیوترنسلیت[1] پرداخته می‌شود. در ارزیابی جعبه‌سیاه نتیجه کلی نظام تولید شده اهمیت دارد و نظام پردازش زبان طبیعی مورد نظر به صورت یکپارچه مورد ارزیابی قرار می‌گیرد. پس از انتخاب داده آزمون، پرس‌وجوها توسط ماشین ترجمه نظام رایسست‌کیوترنسلیت ترجمه شدند. پس از ترجمه، ۵۰ پرس‌وجوی انگلیسی با کمک یک پزشک عمومی با بیش از ده سال سابقه فعالیت، در پایگاه اطلاعاتی مقالات انگلیسی رایسست[2] جستجو شد (انجام قضاوت در خصوص مرتبط بودن یا نبودن اسناد بازیابی شده بصورت انسانی توسط فرد متخصص بررسی شد). پرس‌وجوهای فارسی نیز به همین ترتیب در پایگاه مقالات فارسی رایسست[3] توسط نظام بازیابی اطلاعات بین زبانی رایسست‌کیوترنسلیت به صورت خودکار جستجو و بازیابی گردید. در ادامه به نحوهٔ محاسبه میانگین متوسط دقت پرداخته می‌شود.

جهت محاسبهٔ میانگین متوسط دقت، ابتدا مقدار میانگین دقت[4] برای همهٔ پرس‌وجوهای انگلیسی و فارسی محاسبه گردید. برای محاسبهٔ میانگین دقت برای هر پرس‌وجو، مدل 11 interpolated Average Precision به کار گرفته شد. دلیل به کار گیری این مدل آن بود که به ازاء هر پرس‌وجو، تعداد زیادی رکورد بازیابی می‌شد که محاسبه مقدار میانگین دقت در همهٔ جایگاه‌ها غیر ممکن بود لذا با استفاده از این مدل، مقدار میانگین دقت برای هر

1 RICeSTQUTranslate
2 https://search.ricest.ac.ir/ricest/eearticle.aspx
3 https://search.ricest.ac.ir/ricest/earticle.aspx
4 Average precision

پرس‌وجو تا جایگاه یازدهم (با توجه به مختصات مدل) محاسبه گردید. فرمول زیر مراحل محاسبه میانگین دقت را نشان می‌دهد:

$$P11 - pt = \frac{1}{11}\sum_{j=0}^{10}\frac{1}{N}\sum_{i=1}^{N} pi(rj) \qquad (2)$$

در فرمول (2)، میزان دقت تا جایگاه یازدهم سندهای بازیابی شده محاسبه می‌گردد. rj نقطه استاندارد بازیابی محسوب می‌شود که مقدار آن یازده می‌باشد که شیوهء محاسبه آن در ادامه آمده است. اگر تمام سندهای بازیابی شده مرتبط با پرس‌وجو مورد نظر باشند مقدار میانگین دقت، برابر است با ۱ در غیر این صورت کمتر از ۱ است. Pi در این معادله دائماً تغییر می‌کند چون هر لحظه ممکن است سند جدیدی و مرتبط تا جایگاه یازدهم بازیابی شود. پس‌از حصول تمام نمرات تا جایگاه یازدهم، میانگین این نمرات محاسبه می‌شود و میانگین دقت مربوط به این پرس‌وجو به‌دست می‌آید. در مرحله بعدی میانگین تمام نمرات میانگین دقت مقدار میانگین دقت متوسط را تشکیل می‌دهد. در فرمول زیر نحوه محاسبه‌ی میانگین دقت متوسط بیان می‌شود:

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Qj}\sum_{i=1}^{Qj} p(doc\ i) \qquad (3)$$

در فرمول (۳)، Qj تعداد سندهای بازیابی شده به‌ازا پرس‌وجوی N، j تعداد همه پرس‌وجوها و p(doc i) مقدار دقت در جایگاه iام است.

پس از محاسبهٔ میانگین دقت برای تمام پرس‌وجوها، میانگین این مقادیر مقدار میانگین متوسط دقت را تشکیل می‌دهد. در جدول زیر مقادیر میانگین متوسط دقت برای پرس‌وجوهای انگلیسی و فارسی آورده شده است:

جدول ۴. میزان عملکرد نظام‌های تک زبانه و بین زبانی

نظام	تک زبانه	بین زبانی
میانگین متوسط دقت	۰/۵۴۰	۰/۴۱۰

این مقدار میانگین متوسط دقت با توجه به نامتوازن بودن تعداد مقالات در پایگاه‌های فارسی و انگلیسی رایسست، بدست آمده است. تعداد مقالات در پایگاه انگلیسی 2,392,549 و در پایگاه فارسی 932,395 بوده است. در این پژوهش، تفاوت در تعداد مقالات بر مقدار میانگین متوسط دقت تاثیر منفی گذاشته است.

۳-۲. ارزیابی جعبه‌شیشه‌ای

در این بخش عملکرد ماشین ترجمه به عنوان مهم‌ترین جزء نظام بازیابی اطلاعات بین زبانی رایسست‌کیوترنسلیت مورد ارزیابی قرار گرفت. ارزیابی با استفاده از روش استاندارد بلو[1] [25] صورت پذیرفت. با بکارگیری ۵۰ پرس‌وجوی انتخاب شده در داده آمون، این پرس‌وجوها توسط نظام رایسست‌کیوترنسلیت ترجمه شد. سپس این پرس‌وجوها توسط دو پزشک عمومی نیز ترجمه گردید. پس از آن با استفاده از بسته نرم افزاری برخط آسیا آنلاین[2] آزمایش بلو روی پرس‌وجوهای ترجمه شده صورت گرفت.

در روش بلو با محاسبهء دقت n-gramها در متن ترجمه شده توسط انسان و متن ترجمه شده به وسیلهٔ ماشین ترجمه می‌توان شباهت این دو متن را بدست آورد. هرچه این دو متن شباهت بیشتری داشته باشند دقت ترجمه نیز بالاتر می‌رود. روش محاسبه بلو در ادامه بررسی خواهد شد.

$$pn = \frac{\sum_{i=1}^{1}\sum(n-gram)\epsilon system \;\; count(n-gram)}{\sum_{i=1}^{1}\sum(n-gram)\epsilon reference \;\; count(n-gram)} \qquad (4)$$

در معادله 4، «system» جملات ترجمه شده توسط ماشین و «reference» جملات ترجمه شده توسط انسان هستند. میزان دقت n-gram در جملات ماشین و مرجع انسانی توسط معادله بالا محاسبه می‌شود.

$$BP = Min\left(1, e^{1-\frac{r}{c}}\right) \qquad (5)$$

فرمول 5، مقدار جریمه اختصار گویی را محاسبه می‌کند. این مقدار زمانی به وجود می‌آید که طول جمله ماشین ترجمه کوتاهتر از طول جمله ترجمه شده توسط انسان باشد. در این

1 BLEU
2 http://asiya.lsi.upc.edu/demo/asiya_online.php

فرمول، c طول جمله ترجمه شده توسط ماشین و r طول جمله ترجمه شده توسط انسان است و در نهایت مقدار بلو توسط فرمول زیر محاسبه می‌شود:

$$BLEU = BP * \exp(\sum_{n=1/\frac{1}{n}}^{3} \log(p(n)) \qquad (6)$$

در فرمول 6، n تعداد جملات جهت ترجمه و p(n) مقدار دقت n-gram است که توسط فرمول بالا محاسبه می‌شود. در این پژوهش پس از انجام آزمایش ماشین ترجمه انگلیسی به فارسی به ازاء ۵۰ پرس‌وجوی انگلیسی توسط مدل بلو مورد ارزیابی قرار گرفته که نتایج بدست آمده در جدول زیر آورده شده است:

جدول ۵. میزان آزمایش بلو

ترجمه انگلیسی به فارسی	جملات	نمره بدست آمده
آزمایش بلو	50	0/455

در این آزمایش هر چه نمره کسب شده به عدد صفر نزدیک‌تر باشد خطای ماشین ترجمه کمتر بوده و در نتیجه دقت عملکرد ماشین ترجمه بیشتر است. طراحان مدل بلو [25] ذکر کرده‌اند که طبق این مدل حتی ترجمه‌های انسانی نیز بیش از 0/3 نمره را کسب نکرده‌اند. هرچه عدد بدست آمده به ۱ نزدیک‌تر باشد ترجمه دقیق‌تری انجام گرفته است. در ادامه به بحث و نتیجه گیری در پژوهش حاضر پرداخته خواهد شد.

۴. جمع‌بندی و نتیجه‌گیری

در پژوهش حاضر چند هدف عمده دنبال شد. اولین و مهمترین هدف انجام آن پژوهش این بود که با استفاده از یک نرم‌افزار بازیابی اطلاعات بین زبانی امکانی فراهم گردد تا: (۱) کاربران غیر فارسی زبان نیز از پایگاه مقالات فارسی رایسست استفاده نمایند و مشکل موانع زبانی برای این کاربران حل گردد؛ (۲) جلوی اتلاف زمان کاربران رایسست جهت جستجو در زبان‌های مختلف گرفته شود که این وظیفه را نظام بازیابی اطلاعات رایسست‌کیوترنسلیت بر عهده دارد. همچنین، با استفاده از نظام پیشنهاد پرس‌وجو روال جستجو بسیار تسهیل گردید.

پس از انتخاب دادهٔ پژوهش، نظام بازیابی اطلاعات بین زبانی رایسست‌کیوترنسلیت طراحی گردید. این نرم‌افزار سه جزء اصلی دارد. جزء اول، نظام پیشنهاد پرس‌وجو می‌باشد که فرایند جستجو را به وسیله ارائهٔ پرس‌وجوها تسهیل می‌کند. در قسمت پیشنهاد پرس‌وجو از یک الگوریتم تغییر ترتیب اولویت پیشنهاد جدید استفاده گردید. جزء دوم، که قسمت اصلی نرم‌افزار رایسست‌کیوترنسلیت نیز می‌باشد، وظیفه ترجمه و جستجوی پرس‌وجوها را برعهده دارد. قسمت سوم نرم‌افزار رایسست‌کیوترنسلیت نیز جهت کمک بیشتر به کاربران غیرفارسی زبان جهت ترجمه عناوین اسناد بازیابی شده تعبیه شده است. نرم‌افزار رایسست‌کیوترنسلیت در قسمت مقالات انگلیسی پایگاه رایسست بارگذاری می‌شود. کاربران غیر فارسی زبان با بکارگیری این نرم‌افزار به راحتی و بدون موانع زبانی می‌توانند در پایگاه مقالات فارسی رایسست اطلاعات مورد نیازشان را بازیابی نمایند. این ابزار از اتلاف وقت این کاربران تا حد ممکن جلوگیری می‌کند. با استفاده از این نرم‌افزار هوشمند نیازی نیست تا مطالب در همهٔ زبان‌ها تهیه گردد، از این رو در هزینه‌ها نیز صرفه‌جویی می‌شود.

جهت ارزیابی نظام بازیابی اطلاعات رایسست‌کیوترنسلیت، دو روش معمول ارزیابی نظام‌های بازیابی اطلاعات بین زبانی یعنی ارزیابی جعبه‌سیاه و جعبه‌شیشه‌ای مورد استفاده قرار گرفت. در ارزیابی در مدل جعبه‌سیاه، نتیجهٔ کلی نظام بازیابی اطلاعات بین زبانی رایسست‌کیوترنسلیت مورد ارزیابی قرار گرفت. در مدل جعبه‌شیشه‌ای، الگوریتم ترجمهٔ خودکار پیشنهاد شده طبق معیار بلو و ور ارزیابی گردید.

جهت ارزیابی عملکرد نظام بازیابی اطلاعات بین زبانی رایسست‌کیوترنسلیت، نتایج بدست آمده از این نظام هوشمند با دیگر پژوهش‌های مرتبط مقایسه گردید. همانطور که در پاراگراف قبلی بیان شد، داده ارزیابی مورد استفاده در این پژوهش داده استاندارد ترک بوده است. بدلیل عدم وجود پژوهش انجام شده با رویکرد پژوهش حاضر در حوزه زبان فارسی، لذا برای ارزیابی عملکرد نرم‌افزار طراحی شده در این پژوهش، نتایج بدست آمده با پژوهش‌های انجام شده در دیگر زبان‌ها مقایسه گردید. صالح و پسینا[2] ، عملیات بازیابی اطلاعات بین زبانی را با رویکرد مشابه در پژوهش حاضر بین چندین زبان از جمله انگلیسی و اسپانیایی به انجام رساندند. نتایج این پژوهش در جدول به شرح زیر می‌باشد:

جدول ۶. نتایج پژوهش صالح و پسینا ۲۰۱۶

نظام	تک زبانه	بین زبانی	عملکرد
انگلیسی اسپانیایی	۰/۲۹۹	۰/۲۳۸	**۷۹.۵٪**
انگلیسی مجارستانی	۰/۲۹۹	۰/۲۲۵	**۷۵.۲٪**
انگلیسی لهستانی	۰/۲۹۹	۰/۱۹۷	**۶۵.۹٪**
انگلیسی سوئدی	۰/۲۵۹	۰/۱۹۸	**۷۶٪**

سطر اول جدول ۵، نشان‌دهنده‌ی دو نوع نظام بازیابی اطلاعات تک‌زبانه و بین زبانی می‌باشد. در سطرهای بعدی به ترتیب نتایج بین این نظام‌ها و در زبان‌های مختلف مانند انگلیسی و لهستانی نشان داده شده است. ستون دوم حاوی نمرات کسب شده‌ی آزمایش میانگین متوسط دقت[1] توسط نظام‌های بازیابی اطلاعات تک زبانه و ستون سوم نشان‌دهندی همین نمره توسط نظام‌های بازیابی اطلاعات بین زبانی می‌باشد. در ستون چهارم درصد شباهت عملکرد این دو نظام با یکدیگر نشان داده شده است. در سطر دوم جدول ۵-۱ به ترتیب نظام‌های بازیابی اطلاعات تک‌زبانه و بین زبانی نمایش داده شده است. جهت ارزیابی عملکرد این دو نظام، پژوهشگران از مدل میانگین متوسط دقت استفاده کردند. در ستون دوم نمرات کسب شده توسط نظام‌های بازیابی اطلاعات تک‌زبانه و در ستون سوم نیز نمره کسب شده توسط نظام‌های بازیابی اطلاعات بین زبانی گزارش شده است. و در نهایت در ستون چهارم، درصد شباهت عملکرد ایندو نظام آورده شده است. در سطر دوم، نحوه‌ی عملکرد نظام‌های بازیابی اطلاعات بین زبانی انگلیسی و اسپانیایی نشان داده شده است. عملکرد نظام بازیابی اطلاعات بین زبانی با کسب نمره ۰/۲۳۸ حدود ۷۹/۵ درصد شبیه به نظام بازیابی اطلاعات تک‌زبانه ۰/۲۹۹ گزارش شد. سطرهای بعدی به‌ترتیب، دقت نظام‌های بازیابی اطلاعات بین

1 MAP score

زبانی انگلیسی و مجارستانی، انگلیسی و لهستانی و انگلیسی و سوئدی را با عملکردهای ۷۵/۲ درصد، ۶۵/۹ درصد و ۷۶ درصد شبیه به نظام‌های تک‌زبانه نشان می‌دهد. در پژوهش حاضر نیز از میانگین متوسط دقت جهت ارزیابی دقت نظام بازیابی اطلاعات بین زبانی رایسست‌کیوترنسلیت و نظام تک‌زبانه رایسست استفاده شد. نتایج بدست آمده در پژوهش حاضر به شرح زیر است:

جدول ۷. نتایج شباهت عملکرد نظام تک زبانه و بین زبانی

نظام	تک زبانه	رایسست‌کیوترنسلیت	عملکرد
انگلیسی فارسی	۰/۵۴۴	۰/۴۱۱	۷۵٪

با توجه به نمره میانگین متوسط دقت بدست آمده توسط نظام بین زبانی رایسست‌کیوترنسلیت (0/411) و نظام بازیابی اطلاعات تک زبانه (0/544)، می‌توان نتیجه گرفت که عملکرد رایسست‌کیو ترنسلیت ۷۵ درصد شبیه به نظام بازیابی اطلاعات تک زبانه رایسست بوده است. با مقاسیه عملکرد رایسست‌کیوترنسلیت و نظام‌های موجود در پژوهش صالح و پسینا [2] می‌توان به این نتیجه رسید که عملکرد رایسست‌کیوترنسلیت رضایت‌بخش است. با توجه به نمره کسب شده در ارزیابی ماشین ترجمه، می‌شود نتیجه گرفت که الگوریتم پیشنهادی ترجمهٔ پرس‌وجو در عین سادگی به خوبی عمل نموده است و می‌توان جهت ترجمهٔ پرس‌وجو در فرایند بازیابی اطلاعات بین زبانی، این الگوریتم را با روش‌های دیگر مانند الگوریتم‌های آماری جایگزین نمود.

منابع

[1] Rao, R. & Madankar, M. (Eds.). (2017). Proceedings of IRF International Conference. Bhopal: India.

[2] Saleh, S. & Pecina, P. (Eds.). (2016). Proceedings of the Medical Information Retrieval (MedIR) Workshop. Pisa: Italy.

[3] Iswarya, P. & Radha, V. (2017). Adapting hybrid machine translation techniques for cross-language text retrieval system. Journal of

Engineering Science and Technology, 12(3), 648-666.

[4] Bhattacharya, P., Goyal, P. & Sarkar, S. (Eds.). (2016). Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing. Osaka: Japan.

[5] Sharma, M. & Morwal, S. (2015). A survey on cross language information retrieval. International Journal of Advanced Research in Computer and Communication Engineering, 4(2), 384-387.

[6] Amelina Nasharuddin, N. & Taufik Abdullah, M. (2010). Cross-lingual information retrieval. Electronic Journal of Computer Science and Information Technology (eJCSIT), 2(1), 1-5.

[7] Oard, D. W. (Eds.). (1998). Proceedings from AMTA. Langhorne, PA: USA.

[8] Pothula, S. & Havachelvan, D. (2011). A review on the cross and multilingual information retrieval. International Journal of Web & Semantic Technology (IJWesT), 2(4), 115-124.

[9] Ture, F., Lin, J. & Oard, D. W. (2012). Combining statistical translation techniques for cross-language information retrieval. Proceedings of COLING 2012: Technical Papers, 12, 2685-2702.

[10] Ballesteros, L. & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. ACM, 31, 84-91.

[11] Udupa, R. & Maji, H. K. (Eds.). (2006). Proceedings from the 11st Conference of the European Chapter of the Association for Computational Linguistics. Trento: Italy.

[12] Oard, D. W. (Eds.). (1998). Proceedings from AMTA. Langhorne, PA: USA.

[13] Nikoulina, V., Kovachev, B., Lagos, N. & Monz, C. (Eds.). (2012). Proceedings from the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon: France.

[14] Gupta, S. K., Sinha, A. & Jain, M. (2011). Cross lingual information retrieval with SMT and query mining. Advanced Computing: An International Journal (ACIJ), 2(5), 33-39.

[15] Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, D., Hiemstra, D. & De Jong, F. (Eds.). (2008). Proceedings from CLEF 2008. Aarhus: Denmark.

[16] Rahmani, A., Falahati Fumani Qadimi, M. R. & Dastgheib, M. B. (Eds.). (2017). Proceedings from 19th Artificial Intelligence and Signal Processing Conference (AISP 2017). Shiraz: IR.Iran.

[17] Saravanan, K., Udupa, R. & Kumaran, A. (Eds.). (2013). Proceedings from Improving Cross-Language Information Retrieval by Transliteration Mining and Generation. Berlin: Germany.

[18] Manikandan, B., and Shriram, R. (Eds.). (2011). Proceedings from Third International Conference on Electronics Computer Technology. Kanyakumari: India.

[19] Pemawat, V.; Saund, A.; and Agrawal, A. (2010). Hindi - English based cross language information retrieval system for Allahabad Museum. Proceedings of International Conference on Signal and image processing (ICSIP).

[20] Antony, P.J.; Ajith, V.P.; and Soman, K.P. (2010). Kernel method for English to Kannada transliteration. Proceedings of Recent Trends in Information, Telecommunication and Computing (ITC). Kochi, 336-338.

[21] Chaware, S.M.; and Srikantha, R. (2009). Domain Specific Information Retrieval in Multilingual environment. International journal of recent trends in Engineering and technology, 2(4), 179-181.

[22] Kim, Y., Seo, J., Croft, W. B. & Smith, D. A. (2014). Automatic suggestion of phrasal-concept queries for literature search. Information Processing and Management, 50, 568-583.

[23] Baeza-Yates, R., Hurtado, C. & Mendoza, M. (2004). Query recommendation using query logs in search engines. In Proceedings of the 2004 International Conference on Current Trends in Database Technology (EDBT), 4, 588-596.

[24] Jones, R., Rey, B., Madani, O. & Greiner, W. (2004). Generating query substitutions. In Proceedings of the 15th International Conference on World Wide Web(WWW), 6, 387-396.

[25] Papineni, K., Roukos, S., Ward, T. & Zhu, W. (Eds.). (2002). Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Pennsylvania: USA.

طراحی یک نظام هوشمند جهت بررسی صحت املایی کلمات متون خبری زبان فارسی

امین رحمانی*، صادق خندانی، ایمان میرزاه‌خواه*** و پروانه کهن‌زاد****

چکیده

لزوم صحت املایی و نگارشی محتوای تولید شده در فضای مجازی و غیرمجازی جهت جلوگیری از زوال تدریجی زبان فارسی نقش ویراستاران را به خوبی توجیه می‌نماید. نرم‌افزارهای غلط‌یاب، به عنوان ابزار کمکی، ویراستاری متن را تسهیل و تسریع می‌بخشد.بر همین اساس، هدف از انجام پژوهش حاضر طراحی یک نرم‌افزار جهت بررسی صحت املایی متون خبری زبان فارسی می‌باشد. در این پژوهش، جهت تشخیص غلط‌های احتمالی در متون، از روش جستجو در واژه‌نامه استفاده شده است. در سایر پژوهش‌ها، کلمات متن باکلمات صحیح درون واژه‌نامه مقایسه می‌شوند و اگر کلمه‌ای از متن در واژه‌نامه موجود نباشد، کلمه مورد نظر به‌عنوان ناواژه در نظر گرفته می‌شود، اما در پژوهش حاضر در فرآیند غلط‌یابی، کلمات متن ورودی با کلمات ناواژه در واژه‌نامه مقایسه می‌شوند و سپس ناواژه‌ها در متن مشخص می‌شوند. با به کارگیری روش فاصله ویرایشی نظام پیشنهاد واژه نیز طراحی گردیده است. قسمت غلط‌یابی و همچنین پیشنهاد واژه در نرم‌افزار طراحی شده و نرم‌افزار برخط ویراست‌لایو[1] بر اساس دو معیار دقت و جامعیت مورد بررسی قرار گرفت. بر اساس معیار اف، ماژول‌های غلط‌یابی و تصحیح غلط در نرم‌افزار حاضر به ترتیب عملکرد حدود ۹۲/۵ درصد و ۹۰ درصد را ارائه دادند. ویراست‌لایو نیز با عملکرد ۶۰ درصد و ۶۳ درصد برای هردو ماژول، نسبت به نرم افزار تولید شده از دقت و جامعیت کمتری در

* دانشجوی کارشناسی‌ارشد، زبانشناسی رایانشی، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری شیراز، A.rahmani@ricest.ac.ir

** دانشجوی کارشناسی‌ارشد، زبانشناسی رایانشی، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری شیراز، S.khandani@ricest.ac.ir

*** دانشجوی کارشناسی‌ارشد، زبانشناسی رایانشی، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری شیراز، Iman.khah1993@gmail.com

**** دانشجوی کارشناسی‌ارشد، زبانشناسی رایانشی، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری شیراز، P.kohanzad96@ricest.ac.ir

1 www.virastlive.com

تشخیص و تصحیح غلط‌های موجود برخوردار است. بر اساس نتایج بدست آمده، رویکرد ارائه شده در پژوهش حاضر را می‌توان به شرط وجود یک واژه‌نامه کامل، رویکردی موثر و مفید دانست.

واژه‌های کلیدی: ناواژه، فاصله ویرایشی، جستجو در واژه نامه، دقت، جامعیت، معیار اف.

۱. مقدمه

با گسترش فضای اینترنتی و تولید محتوا، خصوصا محتوای متنی، لزوم صحیح بودن متون تولید شده از اهمیت ویژه‌ای برخوردار است. خبر، به‌عنوان یکی از مهمترین تولیدات در فضای مجازی بیشترین سهم از حجم محتویات تولید شده را به خود اختصاص داده است، لذا ویرایش این متون الکترونیکی تولید شده به روش انسانی بسیار وقت‌گیر خواهد بود. ابزارهای هوشمندی مثل غلط‌یاب‌ها می‌تواند به ویراستاران در این زمینه کمک شایان بنماید[1]. اولین غلط‌یاب هوشمند برای زبان انگلیسی در سال ۱۹۸۰ تولید شد[2]. سپس پلک و زامورا[1] [3] با استفاده از یک واژه‌نامه غلط‌یاب جدیدی را طراحی نمودند. اتول و الیوت[2] [4] نیز با استفاده از روش ان-گرم[3] غلط‌یابی را جهت تشخیص اشتباهات موجود در متن طراحی کردند. مانگو و بریل[4] [5] روش جدیدی را جهت پیشنهاد کلمه صحیح ارائه کردند. این روش بر اساس تغییر جایگاه نویسه‌های کلمه طراحی گردید.

در ادامه این بخش به بررسی روش‌های غلط‌یابی و تصحیح این غلط‌ها پرداخته خواهد شد. ابتدا انواع غلط‌های املایی در متون بررسی می‌شود. جدول ۱ نمونه‌ای از غلط‌های املایی را نشان می‌دهد:

1 Pollock and Zamora
2 Atwell and Elliott
3 N-gram
4 Mangu and Brill

جدول ۱. نمونه غلط‌های املایی

کلمه غلط	کلمه صحیح	نوع خطا
آشسکارتر	آشکارتر	درج
آدس	آدرس	حذف
آاهی	آگاهی	جایگزینی
آلمنا	آلمان	جابجایی

در جدول ۱ نمونه غلط‌های معمول املایی قابل مشاهده است. در این نوع غلط‌ها چهار حالت عمده درج نویسه در کلمه، حذف نویسه از کلمه، جایگزینی نویسه به جای نویسه دیگر و در نهایت جابجایی نویسه‌های هم‌جوار اتفاق می‌افتد. در ادامه روش‌های غلط‌یابی و تصحیح این نوع غلط‌ها مورد بررسی قرار می‌گیرد.

از مهمترین روش‌های تشخیص خطاهای املایی می‌توان به روش اِن-گرم و روش مبتنی‌بر واژه‌نامه اشاره کرد[۶]. در روش اِن-گرم مجموعه‌ای از حروف متوالی یک رشته به طول n درنظر گرفته می‌شود. اِن-گرم‌های یک‌حرفی را یونی-گرم، دو حرفی را بای-گرم و سه‌حرفی را ترای-گرم می‌نامند. هر اِن-گرم یک رشته‌ی ورودی می‌باشد که با جدولی از اِن-گرم‌های صحیح که از قبل آماده شده مقایسه می‌شود. در صورت عدم وجود یا رخداد پایین رشته‌ی ورودی سیستم آن را خطا تشخیص می‌دهد. روش مبتنی‌بر واژه‌نامه، هر واژه‌ی ورودی را در واژه‌نامه جست‌وجو می‌کند. اگر واژه در واژه‌نامه موجود باشد آن را صحیح تشخیص می‌دهد در غیر اینصورت آن را در فهرست واژه‌های نادرست ذخیره می‌کند[8,7]. در ادامه روش‌های تصحیح خطاهای املایی مورد بررسی می‌شود.

روش حداقل فاصله ویرایشی [1]: حداقل فاصله ویرایشی یکی از ساده‌ترین روش‌های غلط یابی محسوب می‌شود که غلط یابی را بر مبنای کمترین خطای کاربر انجام می‌دهد[9]. بنابراین برای هر واژه ابتدایی ترین عملیات ویرایشی (درج، حذف، جایگزینی) را که منجر به تبدیل واژه‌ها به ناواژه‌ها می‌شود را در نظر می‌گیرد. :

1 Minimum edit distance

جدول۲. غلط‌های نگارشی همراه با فاصله ویرایشی

فاصله ویرایشی	غلط	صحیح	نوع خطا
۱	آشسکارتر	آشکارتر	درج
۱	آدس	آدرس	حذف
۱	آاهی	آگاهی	جایگزینی
۱	آلمنا	آلمان	جابجایی

در جدول ۲ فاصله ویرایشی مربوط به هر چهار نوع رایج غلط‌های املایی نیز آورده شده است. برای مثال در عملیات درج، میان کلمات آشکارتر و آشسکارتر یک نویسه اضافی وارد شده است که فاصله ویرایشی بین این دو کلمه برابر با یک می‌باشد . لونشتاین[1] نیز در از روشی مشابه همین روش استفاده کرد با این تفاوت که او عملیات ویرایشی درج و حذف و جابجایی را در مدل خود به کار برد[10].

روش کلید شباهت[2]: روش کلید شباهت برای هر واژه و ناواژه کلید تعیین می‌شوند به این ترتیب واژه‌هایی که کلید آن‌ها بیشترین شباهت را با کلید ناواژه‌ها دارند به عنوان پیشنهاد ارائه می‌شوند[10]. این روش در تسریع پردازش تاثیر بسزایی دارد که خود یک مزیت محسوب می‌شود.

روش مبتنی‌بر قاعده : روش مبتنی‌بر قاعده شامل الگوریتم‌هایی است که بر اساس خطاهای رایج املایی طراحی شده و به شکل قاعده درآمده‌اند. این الگوریتم‌ها واژه‌های نادرست را به واژه‌های صحیح تبدیل می‌کنند[12].

روش‌های احتمالی : روش‌های احتمالی مبتنی‌بر ویژگی‌های آماری زبان می‌باشند. این روش به دو رویکرد احتمال جابه‌جایی و احتمال اشتباه تقسیم می‌شود. احتمال جابه‌جایی به روش اِن-گرم شباهت دارد. این روش احتمال رخداد هر حرف پس از حرف دیگر را تخمین می‌زند. رویکرد دیگر احتمال اشتباه می‌باشد که احتمال رخداد یک حرف بجای حرف دیگر را محاسبه می‌کند[12,11].

روش مبتنی‌بر اِن-گرم: اِن-گرم در غلط یابی به دو صورت با استفاده از واژه‌نامه و بدون

1 Levenshtein
2 Similarity key technique

استفاده از آن بکار می‌رود. در صورت عدم وجود واژه‌نامه، مـی‌تـوان بـا اسـتفاده از اِن-گـرم آن قسمت از واژه که در آن خطای املایی رخ داده را پیدا کـرد. در صـورت امکـان تبـدیل واژه ی نادرست به اِن-گرم‌های صحیح می‌توان آن را تغییر داد و به عنوان واژه‌ی صحیح معرفـی کـرد. در صورت وجود واژه‌نامه اِن-گرم‌ها برای تعریف فاصله‌ی میان واژه‌ها بکار خواهند رفت و واژه‌ها دائما با واژه‌نامه تطبیق داده خواهند شد. بدین ترتیب که اِن-گـرم‌هـای واژه‌ی نادرسـت بـا اِن-گرم‌های واژه‌ی درون واژه‌نامه با یکدیگر مقایسه می‌شوند[8][12].

۱-۱. پژوهش‌های انجام شده در زبان‌های خارجی

الگوریتم‌های متافون [۱] و متافون دوتایی هر کاراکتر از الفبای زبان را کنتـرل کـرده و روی تلفظ هر هجا از کلمه متمرکز می‌شود. از آنجایی که زبان‌های غربی الفبای محدودتری دارنـد و از پیچیدگی کمتری برخوردار هستند، این الگوریتم‌ها برای شناسـایی پیشـنهادها در ایـن نـوع زبان‌ها مناسب می‌باشند. زبان‌های آسیای جنوبی جز آن دسته از زبـان‌هـا هسـتند کـه الفبـای آن‌ها گسترده و پیچیده می‌باشد. بنابراین در این نوع زبان‌ها نمی‌توان از الگوریتم‌های متـافون و متافون دوتایی بهره گرفت. برای ارائه پیشنهاد مناسب در اینگونه زبان‌ها، الگوریتمی نیـاز اسـت تا واژه‌های غلط را بر اساس مجموعه حروف مشـابه شـبیه‌سـازی کنـد. شـبیه‌سـازی بازگشـتی [۲]الگوریتمی است که غلط یابی در این نوع زبان‌ها را تا حدی پوشش می‌دهد[13].

زبان عربی جز آن دسته از زبان‌ها می‌باشد که ساخت‌واژه غنی و پیچیده دارند. غلط یـابی این نوع زبان‌ها چالش‌های خاص خود را دارد. غلط یابی در زبان‌های تصریفی مانند زبـان عربـی نیازمند روشی مبتنی بر جست‌وجوی واژه‌نامه[۳] و تحلیل ساخت‌واژه‌ای می‌باشد. الگـوریتم غلـط یاب در این نوع زبان‌ها را می‌توان با استفاده از این روش‌ها بهبود بخشـید. پژوهشـگرانی ماننـد شالن[۴]، الم[۵] و گومه[۶][14] با تهیه لیستی از واژه‌های صحیح و غلط روشی مبتنی بر ترای-گـرم[۷] حروف در مدل زبان را به وجود آوردند. با این روش دانش حـروف مجـاز در عربـی تخمـین زده

1 metaphone
2 Recursive Simulation
3 lexicon
4 Khaled Shaalan
5 Amin Allam
6 AbdAllah Gomah
7 trigram

می‌شود و خطاهای املایی شناسایی می‌شوند.

غلط یابی در برخی زبان‌ها نیازمند الگوریتم‌های ترکیبی و چندمرحله‌ای می‌باشد. برای نمونه می‌توان به پژوهشی که بیک[1] [15] در زبان دانمارکی انجام داده است اشاره کرد. او در مدل خود از لیست خطاهای داده‌محور[2]، میزان شباهت نظام آوایی و تطابق حرف در مرحله‌ی واژه و قطعه[3] استفاده کرده است و همچنین در مرحله‌ی بافت constraint grammar را در مدل خود به کار برده است. این روش موفق شده ۷۰ درصد خطاها را تشخیص دهد و مقدار تقریبی نمره اف آن ۴۴ بدست آمده است.

از مدل‌های دیگر در غلط یابی می‌توان به مدل کانال‌های نویزی[4] اشاره کرد. اینگونه مدل‌ها از دو بخش مدل منبع[5] و مدل کانال[6] تشکیل شده‌اند. بریل[7] و مور[8] [16] با بهبود مدل کانال مدلی جدید در غلط یابی ارائه کردند. به ادعای این دو نفر این مدل عملکرد مدل‌های قبلی را بهبود بخشیده و خطاهای املایی را ۷۴٪ کاهش داده است. در جدول ۳ خلاصه‌ای از جدیدترین پژوهش‌های انجام شده در حوزه غلط‌یاب‌ها ارائه شده است:

جدول ۳. خلاصه‌ای از پژوهش‌های انجام شده در حوزه غلطیاب ها

سال	نویسندگان	مدل	نتیجه
۲۰۱۷	Bhirud Bhavsar and Pawar [۱۱]	یک پژوهش مروری	
۲۰۱۷	Ayegba, Ugbedeojo, Jessica Chinezie and Abu [۱۷]	جستجو در واژه نامه، روش کلید شباهت	توانایی صحیح کردن انواع غلط‌های املایی زبان ایگالا
۲۰۱۷	Mandal and Hossain [۱۸]	روش خوشه بندی	۹۹/۸ درصد دقت در تصحیح غلط ها
۲۰۱۷	Alva and Marcos[۱۹]	روش مبتنی بر قانون	ارائه یک مجموعه داده جدید

1 Eckhard Bick
2 data-driven
3 chunk
4 noisy channel
5 source model
6 Channel model
7 Brill
8 Moore

نتیجه	مدل	نویسندگان	سال
افزایش دقت	مدل کانال نویزی	Sorokin[۲۰]	۲۰۱۷
ارائه یک رویکرد جدید و یک واژه نامه جدید	جستجو در واژه نامه، روش آماری	Sarma Goswami2 and Goswami [21]	۲۰۱۶

در جدول ۳ جدیدترین پژوهش‌ها در حوزه غلط‌یابی و تصحیح غلط‌های املایی گنجانده شده است. این پژوهش‌ها عمدتا در سال‌های ۲۰۱۷ انجام گرفته‌اند. در ادامه مباحث، پژوهش‌های انجام شده در حوزه زبان فارسی مورد بررسی قرار خواهد گرفت.

۱-۲. پژوهش‌های انجام شده در زمینه زبان فارسی

علیرغم وجود پژوهش‌های انجام شده در حوزه‌ی غلط‌یابی کلمات در زبان فارسی، نیاز است که پژوهش‌های بیشتری در این زمینه انجام بپذیرد. از جمله اثرات ارائه شده در زبان فارسی می‌توان به پژوهش موسوی [22] اشاره کرد. پژوهشگر برای غلط‌یابی متون از روش جستجو در واژه‌نامه استفاده نمود. در این پژوهش از یک مدل ترکیبی برای ارائه پیشنهاد کلمات صحیح به کاربران بهره گرفت شده است. محقق جهت تصحیح غلط‌های موجود در متن، از یک پیکره استفاده نمود. در قسمت ارائه پیشنهاد به کاربر، از روش فاصله ویرایشی و فراوانی نسبی کلمات در پیکره به صورت ترکیبی استفاده شده است. دقت تشخیص این غلط‌یاب ۹۶ درصد گزارش شد. همچنین دقت نظام رتبه‌بندی پیشنهاد در این سامانه ۹۵ درصد گزارش گردید.

غلط‌یاب وفا [23] از جمله پژوهش‌های دیگری است که در زمینه زبان فارسی انجام پذیرفته است. در این پژوهش با ترکیب اصول مدل آماری و مدل مبتنی بر قانون، از یک روش ترکیبی جهت غلط‌یابی و تصحیح کلمات در متن استفاده شده است. پس از ارزیابی نظام تولید شده بر اساس نمره اف، دقت این نظام در تشخیص کلمات ناواژه حدود ۹۰ درصد گزارش شد.

دستغیب و همکاران [24] جهت تشخیص غلط‌های موجود در متن از روش معنایی بهره جستند. پس از ارزیابی نظام تولید شده تحت عنوان پرسپل، دقت این غلط‌یاب معنایی ۹۸ درصد گزارش گردید. این نظام در مقایسه با نمونه‌های مشابه مانند ویراستیار و وفا از دقت قابل توجه‌تری برخوردار است. همچنین دقت این غلط‌یاب برای ارائه پیشنهاد به کاربر حدود ۸۸

درصد گزارش شده است.

علیرغم وجود نرم‌افزارهای تولید شده مانند ویراستیار، وفا و پرسپل در زمینه غلط‌یابی و اصلاح متون زبان فارسی، دقت این‌گونه نرم‌افزارها در مقایسه با نمونه‌های خارجی کمتر است پس می‌توان از پیشینه پژوهش نتیجه گرفت، جهت تولید یک نرم‌افزار جامع و بادقت نیاز است تا پژوهش‌های بیشتری در این زمینه، در زبان فارسی، صورت بگیرد. . در ادامه تمایزات پژوهش حاضر با پژوهش‌های پیشین بررسی خواهد شد.

در وهله اول (۱) می‌توان به رویکرد جدید غلط‌یابی در پژوهش حاضر اشاره نمود. در پژوهش‌های پیشین، جهت غلط‌یابی، تمام کلمات متن ورودی با کلمات صحیح موجود در واژه‌نامه مقایسه می‌شد و در صورت عدم وجود هر واژه از متن در واژه‌نامه، واژه مذکور به‌عنوان ناواژه در نظر گرفته می‌شد. در پژوهش حاضر این عملیات کاملا بالعکس انجام می‌پذیرد و کلماتی که در واژنامه به عنوان ناواژه مشخص شده‌اند، در متن جستجو و مشخص می‌شوند. در وهله دوم (۲) می‌توان به این نکته اشاره کرد که برای رسیدن به هدف غایی، یعنی تولید یک نرم‌افزار جامع و بادقت، باید به حوزه‌های تخصصی روی آورد. مثلا برای تولید یک غلط‌یاب در حوزه متون علمی و خبری باید پژوهش‌های جداگانه صورت بگیرد. این پژوهش بر پایه‌ی این اصل، در حوزه‌ی غلط‌یابی متون خبری انجام گرفته است. و در درجه سوم (۳) می‌توان به سهولت در استفاده از نرم‌افزار، یعنی رابط کاربری ساده در طراحی نرم‌افزار تولید شده در پژوهش حاضر اشاره نمود. در مقایسه، ویراستیار و نرم‌افزار طراحی شده در پژوهش حاضر هر دو بدون دسترسی به اینترنت نیز قابل استفاده هستند، اما ویراستلایو به‌عنوان یک محصول تجاری بدون وجود اینترنت قابل دسترسی نمی‌باشد. در زمینه غلط‌یابی و پیشنهاد کلمات، نرم‌افزار حاضر با زدن تنها یک کلید غلط‌های مورد نظر را در متن شناسایی می‌کند و سپس با زدن کلید راست موس بر روی کلمات غلط شناسایی شده کلمات صحیح مورد نظر پیشنهاد داده می‌شود اما در ویراستیار باید ناواژه‌ها را تک به تک بررسی کرد که از لحاظ زمانی وقت کاربر بسیار تلف می‌شود. در ادامه بخش‌های دیگر این پژوهش مورد بررسی قرار می‌گیرد.

۲. روش تحقیق

در این بخش ابتدا، داده آموزش در پژوهش حاضر مورد بررسی قرار می‌گیرد. سپس،

الگوریتم و نحوه ساخت نرم‌افزار توضیح داده خواهد شد.

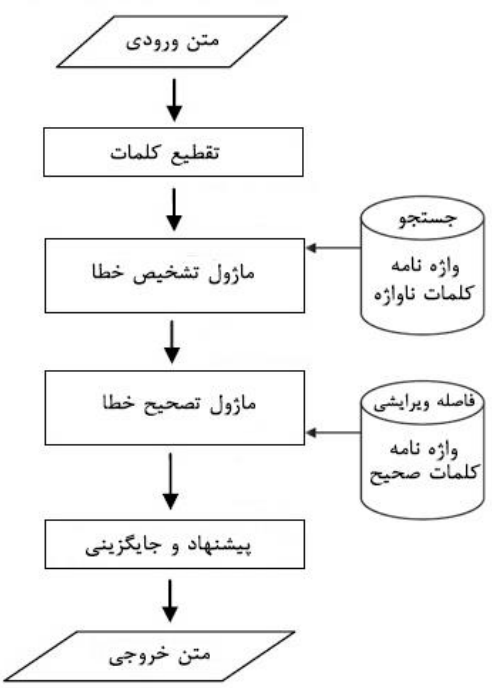
۲-۱. داده پژوهش

جهت آموزش نظام غلط یاب، داده فااسپل [25] مورد استفاده قرار گرفت. این مجموعه، حدودا، شامل پنج‌هزار کلمه غلط و شکل صحیح این کلمات می‌باشد. غلط‌های نگارشی در این مجموعه داده از متون تایپ شده توسط تایپیست‌های حرفه‌ای و متون املایی دانش‌آموزان دوره ابتدایی استخراج گردیده است. همچنین، از مجموع داده دیگری که شامل غلط‌های نگارشی در متون خبری می‌باشد[24]، استفاده گردید. جهت بخش نرمال‌سازی متون نیز افعال و اسامی مرکب پربسامد از حدود دویست مقاله فارسی به‌صورت انسانی ،توسط پژوهشگران، استخراج گردید.

۲-۲. الگوریتم پیشنهادی غلط‌یابی و تصحیح کلمات

در این بخش الگوریتم غلط‌یابی متن مورد بررسی قرار می‌گیرد. این الگوریتم شامل دو مرحله مهم می‌باشد. در مرحله اول، غلط‌های موجود با روش جستجو[1] در لغت‌نامه[2] تشخیص داده می‌شود، سپس با استفاده از روش فاصله ویرایشی[3]، شکل صحیح کلمات به کاربر پیشنهاد می‌گردد. در شکل ۱ الگوریتم غلط‌یابی متون و تصحیح این غلط‌ها در پژوهش حاضر قابل مشاهده است:

1 Look up
2 Lexicon
3 Edit distance

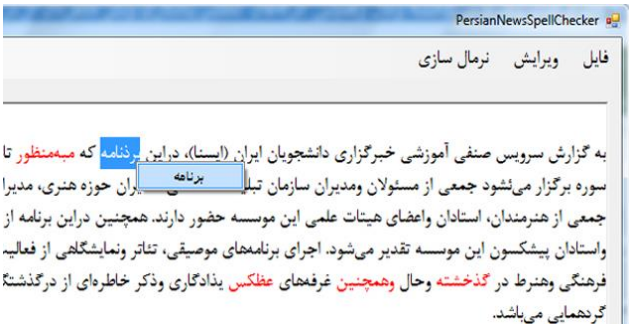


شکل ۱: الگوریتم پیشنهادی تشخیص و تصحیح خطا

با درونداد متن در نرم‌افزار، متن مورد نظر به کلمات تشکیل‌دهنده تقطیع[1] می‌گردد. سپس کلمات تقطیع شده به قسمت تشخیص خطا[2] فرستاده می‌شود. در قسمت تشخیص خطا، تک‌تک کلمات در یک لغت‌نامه جستجو می‌شود. این لغت‌نامه حاوی شکل غلط کلمات می‌باشد. اگر از مجموع کلمات تقطیع‌شده، کلمه‌ای در لغت‌نامه کلمات غلط تطابق یابد، این واژه به عنوان کلمه غلط در متن اصلی به رنگ قرمز نمایان می‌شود

پس از غلط‌یابی متن، با زدن کلید راست موس بر روی کلمات غلط، شکل صحیح کلمه مورد نظر به کاربر پیشنهاد داده می‌شود. در شکل ۲ نحوه‌ی پیشنهاد کلمات صحیح نمایش داده شده‌است:

1 Tokenize
2 Detection module



شکل ۲: نمایی از پیشنهاد کلمه صحیح

شکل صحیح کلمات غلط با استفاده از روش فاصله ویرایشی در لغت‌نامه کلمات صحیح جستجو می‌گردد و سپس به کاربر پیشنهاد داده می‌شود. کاربر قادر است از میان این پیشنهادات شکل صحیح کلمه را انتخاب و در متن جایگذاری نماید.

۳. نتایج

جهت ارزیابی نظام طراحی شده، از یک مجموعه داده آزمایش استفاده شد. این مجموعه شامل هزار تکه متن در حوزه خبر می‌باشد[24]. این متون حاوی غلط‌های نگارشی و املایی می‌باشد. از این هزار تکه متن، پنجاه تکه متن به صورت تصادفی جهت آزمودن نظام طراحی شده انتخاب گردید. جزئیات مربوط به این مجموع داده در جدول ۴ قابل مشاهده می‌باشد:

جدول ۴. جزئیات داده آزمایش

تعداد غلط‌ها	تعداد کلمات متن	تعداد متن
۱۱۳۶	۶۱۹۴	۵۰

پس از تعیین داده آزمایش، همهٔ پنجاه تکه متن در نرم‌افزار درونداد گردید. سپس عملیات غلط‌یابی و صحیح‌سازی کلمات توسط نرم‌افزار انجام پذیرفت. در پژوهش حاضر جهت ارزیابی نظام غلط‌یاب و نظام پیشنهاد کلمه از دو معیار معمول دقت[1] و جامعیت[1] در غلط‌یابی و اصلاح

Precision

کلمات استفاده گردید. پس از حصول نمرات دو معیار مذکور جهت ارزیابی کلی نرم‌افزار از معیار اف[۲] با مقدار آلفا ۰/۵ استفاده گردید. در ادامه به بررسی معیارهای دقت، جامعیت و اف پرداخته خواهد شد. جدول ۵ نحوه بدست آمدن این معیارها را نشان می‌دهد:

جدول ۵. نحوه محاسبه معیار دقت و جامعیت

	صحیح	غیرصحیح	تعداد کل
تشخیص داده شده	A	B	A+B
تشخیص داده نشده	C	D	C+D
تعداد کل	A+C	B+D	A+C+B+D

در این جدول، A کلمات غلطی هستند که به صورت صحیح تشخیص داده شده است. B، کلمات صحیح است که به اشتباه، غلط تشخیص داده شده است و C تعداد کل غلط‌ها در متن می‌باشد.

نحوه مجاسبه دقت در تشخیص و تصحیح کلمات در فرمول ۱ آورده شده است

$$A/(A+B) \quad (۱)$$

میزان جامعیت در تشخیص و اصلاح کلمات نیز در فرمول ۲ بیان شده است:

$$A/C \quad (۲)$$

با توجه به روابط ذکر شده، میزان میانگین دقت و جامعیت در نظام طراحی شده و نظام ویراست‌لایو به شرح زیر می‌باشد:

1 Recall
2 F-Measure

جدول ۶. میانگین دقت و جامعیت ماژول تشخیص و اصلاح خطا

ردیف	نظام	تشخیص		اصلاح	
		دقت	جامعیت	دقت	جامعیت
۱	پژوهش حاضر	۱	۰/۸۶	۰/۹۴	۰/۸۶
۲	ویراست لایو	۰/۸۱	۰/۴۹	۰/۹۰	۰/۴۹

در جدول ۶ ماژول‌های تشخیص و اصلاح نرم‌افزار در پژوهش حاضر و ویراست‌لایو با استفاده از دو معیار دقت و جامعیت ارزیابی گردید. برای محاسبه میزان عملکرد این دو نرم‌افزار معیار اف به کار رفت. فرمول ۳ نحوه محاسبه معیار اف را نشان می‌دهد:

$$f - measure = \frac{2(PR)}{P+R} \qquad (۳)$$

در این فرمول P میزان دقت و R میزان جامعیت است. جدول ۷ نشان دهنده‌ی میزان عملکرد ماژول‌های تشخیص و تصحیح خطا در هر دو نرم‌افزار می‌باشد:

جدول ۷. میزان عملکرد دو نظام بر اساس معیار اف

ردیف	نظام	تشخیص	اصلاح
		معیار اف	معیار اف
۱	پژوهش حاضر	۰/۹۲/۴	۰/۸۹
۲	ویراست لایو	۰/۶۰	۰/۶۳

طبق نتایج بدست آمده، نرم‌افزار طراحی شده در این پژوهش حدود ۹۲/۵ درصد ناواژه‌ها را در داده آزمایش تشخیص داد. از این تعداد غلط، حدود ۹۰ درصد از این غلط‌ها توسط نرم‌افزار مذکور تصحیح شد. این در حالی است که ویراست‌لایو تنها ۶۰ درصد از ناواژه‌های درون داده آزمایش را تشخیص داد و از این تعداد تنها ۶۳ درصد از این غلط‌ها را صحیح نمود. در مقایسه این دو نظام، غلطیاب طراحی شده به مراتب عملکرد بهتری از خود نشان داد. در بخش بعدی به نتیجه گیری خواهیم پرداخت.

۴. جمع‌بندی و نتیجه‌گیری

در پژوهش حاضر سعی بر آن بود تا یک نرم‌افزار با رویکرد جدید جهت بررسی صحت کلمات متون خبری ارائه شود. جهت تشخیص ناواژه‌ها از روش جستجو در واژه‌نامه استفاده گردید. در این پژوهش برخلاف سایر پژوهش‌های مرتبط، برای تشخیص ناواژه‌ها، فهرستی از کلمات غلط در متن جستجو گردید. مزیت این روش سرعت بیشتر در غلط‌یابی متون می‌باشد.جهت ارائه پیشنهاد کلمات صحیح، از روش فاصله ویرایشی بهره گرفته شد. پس از طراحی، ماژول‌های تشخیص و تصحیح خطا، با استفاده از دو معیار دقت و جامعیت مورد ارزیابی قرار گرفت. سپس جهت عملکرد کلی دو ماژول مذکور، با به کارگیری معیار اف نتیجه کلی عملکرد نرم‌افزار طراحی شده نیز محاسبه گردید. نرم‌افزار فوق، قادر بود حدود ۹۳ درصد از ناواژه‌های موجود در داده آزمایش تشخیص دهد و سپس حدود ۹۰ درصد از این غلط‌ها را تصحیح کند. جهت مقایسه عملکرد نرم‌افزار طراحی شده با نرم‌افزارهای موجود، داده آزمایش نیز به وسیله نرم‌افزار برخط ویراست‌لایو مورد بررسی قرار گرفت. پس از محاسبه معیار اف برای نمرات کسب شده توسط ویراست‌لایو، عملکرد نظام ویراست‌لایو در تشخیص غلط ۶۰ درصد و در تصحیح این غلط‌ها ۶۳ درصد بوده است. طبق نتایج بدست‌آمده نرم‌افزار طراحی‌شده در این پژوهش عملکرد چشم‌گیری نسبت به نرم‌افزار ویراست‌لایو از خود نشان داد. در آخر می‌توان نتیجه گرفت که رویکرد ارائه‌شده در این پژوهش بسیار مفید و موثر بوده است. ذکر این نکته ضروری است که پیش‌شرط لازم برای استفاده از رویکرد ارائه شده در پژوهش حاضر، آماده‌سازی یک واژه‌نامه جامع می‌باشد.

منابع

[2] Leacock, C., Chodorow, M. & Gamon, M. (2010). Automated grammatical error detection for language learners. Synthesis Lectures on Human Language Technologies, 3(1), 1-134.

[3] Peterson, J. L. (1980). Computer programs for detecting and correcting spelling errors. Communications of the ACM, 23, 676-687.

[4] Pollock, J. J. & Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text. Communications of the ACM, 27, 358-

368.

[5] Atwell, E. & Elliott, S. (Eds.). (1987). Proceedings from The Computational Analysis of English. London: England.

[6] Mangu, L. & Brill, E. (Eds.). (1997). Proceedings from Proceeding of the 14thInternational Conference on Machine Learning. San Francisco: USA.

[7] Kukich, K. (1992). Techniques for automatically correctingwords in text. ACM Computing Surveys(CSUR), 24(4): 377–439.

[8] Ahmed, F., Luca, W. & Nürnberger, A. (Eds.). (2007). Proceedings from 8th International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City: Mexico.

[9] Wasala, A., Weerasinghe, R. & Gamage, K. (Eds.). (2006). Proceedings from the COLING/ACL Main Conference Poster Sessions. Sydney: Australia.

[10] Damerau, F. (1964). A technique for computer detection and correction of spelling errors. Comm. ACM, 7(3), 171-176.

[11] Wasala, A., Weerasinghe, R., Pushpananda, R., Liyanage, C. & Jayalatharachchi, E. (2010). A Data-Driven Approach to Checking and Correcting Spelling Errors in Sinhala. The International Journal on Advances in ICT for Emerging Regions, 3, 11-24.

[12] Bhirud, N. S., Bhavsar, R. & Pawar, B. (2017). Grammar chekcers for natural languages: A review. International Journal on Natural Language Computing (IJNLC), 6(4), 1-13.

[13] Zhao, H., Cai, D., Xin, Y., Wang, Y. & Jia, Z. (2017). A hybrid model for chinese spelling check. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 16(3), 34-56.

[14] Abdullah, A. B. A. & Rahman, A. (Eds.). (2003). Proceedings from the Second IASTED International Conference on Information and Knowledge Sharing. Scottsdale: USA.

[15] Shaalan, K., Allam, A. & Gomah, A. (Eds.). (2003). Proceedings from the 4th Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE). Cairo: Egypt.

[16] Bick, E. (2006). A constraint grammar based spellchecker for danish with a special focus on dyslexics. SKY Jounal of Linguistics, 19,

387-396.

[17] Brill, E. & Moore, R. (Eds.). (2000). Proceedings from the ACL. Hong Kong: China.

[18] Ayegba, S. F., Ugbedeojo, M., Jessica Chinezie, B. & Abu, O. (2017). Igala language spell checker. Current Journal of Applied Science and Technology, 23(2), 1-9.

[19] Mandal, P. & Hossain, M. M. (Eds.). (2017). Proceedings from 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR). Dhaka: Bangladesh.

[20] Alva, C. & Marcos, A. (Eds.). (2017). Proceedings from the First Workshop on Subword and Character Level Models in NLP. Copenhagen: Denmark.

[21] Sorokin, A. (Eds.). (2017). Proceedings from the 6th Workshop on Balto-Slavic Natural Language Processing. Valencia: Spain.

[22] Sarma, B., Goswami, D. & Goswami, G. (2017). Assamese spell checker design and implementation. International Journal of Modern Trends in Engineering and Research, 3(2), 44-47.

[23] Mosavi Miangah, T. (2013). FarsiSpell: A spell-checking system for Persian using a large monolingual corpus. Literary and Linguistic Computing, 29(1), 55-73.

[24] Faili, H., Ehsan, N., Montazery, M., and Pilehvar, M. T. (2014). Vafa spell-checker for detecting spelling, grammatical, and real-word errors of Persian language. Literary and Linguistic Computing, 28-43.

[25] Dastgheib, M. B., Fakhrahmad, S. M. & Zolghadri Jahromi, M. (2017). Perspell: A new Persian semantic-based spelling correction system. Digital Scholarship in the Humanities, 32(3), 543-553.

[25] Barari, L. & Qasemizadeh, B. (Eds.). (2005). Proceedings from AIML 2005 Conference CICC. Cairo: Egypt.

تحلیل سوگیری زبانی در متون خبری فارسی با روش‌های رایانشی

محدثه عباس‌زاده هجدکی* و **محمد بحرانی****

چکیده

«سوگیری رسانه‌ای» از زوایای گوناگونی قابل بررسی است. این پژوهش به بررسی یک شکل بروز آن یعنی «سوگیری زبانی» در اخبار نوشتاری چهار بنگاه خبری بین‌المللی دویچه‌وله، رادیو بین‌المللی فرانسه، اسپوتنیک و العربیه می‌پردازد. در این راستا متن‌های خبری فارسی مرتبط با ایران از وب‌سایت هر یک از این بنگاه‌ها در دو بازه زمانی قبل و بعد از تصویب برجام جمع‌آوری شده‌اند که مجموعه‌ای بالغ بر ۷۸۴ خبر و ۳۷۶۰۰۰ کلمه را تشکیل می‌دهد. برای تحلیل سوگیری زبانی، ابتدا پیکره‌ای از واژگان سوگیرانه فارسی آماده‌شده‌است سپس هشت ویژگی زبانی و متنی تعریف شده و بر اساس آن‌ها مدلی احتمالاتی به روش بانظارت آموزش‌داده‌شده‌است که دقت و بازخوانی آن در تعیین کلمات سوگیرانه متن به‌ترتیب ٪۷۶ و ٪۷۷ می‌باشد. پس از دسته‌بندی کلمات به سوگیرانه و غیرسوگیرانه و محاسبه نرخ‌های سوگیری زبانی مشخص شد که سوگیری زبانی هر چهار بنگاه خبری مذکور در بازه زمانی پس از برجام نسبت به بازه زمانی پیش از آن افزایش داشته‌است.

واژه‌های کلیدی: سوگیری رسانه‌ای، سوگیری زبانی، واژگان سوگیرانه فارسی

۱ مقدمه

صورت‌های بروز سوگیری رسانه‌ای متنوعند و در قالب‌های گوناگونی مانند انتخاب سوگیرانه موضوعات، استفاده از «زبان سوگیرانه»[1]، منطبق نبودن تیتر با محتوا، تکرار خبر، استفاده از آمار و ارقام و ... ظاهرمی‌شوند؛ با این حال می‌توان با استفاده از روش‌های «پردازش

* کارشناسی ارشد زبان‌شناسی رایانشی دانشگاه صنعتی شریف، abbaszadehcs@gmail.com

** عضو هیات علمی دانشگاه صنعتی شریف، bahrani@sharif.edu

1 Biased language

زبان طبیعی»[1] و انجام فعالیت‌هایی نظیر دسته‌بندی[2] یا خوشه‌بندی[3] خبرها، مشخص‌کردن کلمات سوگیرانه متن، استخراج کلمات کلیدی[4] متن و یا تعیین درصد وزنی هر موضوع در خبرها، رخداد انواعی از سوگیری را شناسایی و تحلیل نمود.

۱-۱ تعریف مسئله

از بین انواع نمودهای سوگیری رسانه‌ای، هدف اصلی پژوهش حاضر، بررسی و تفسیر «سوگیری زبانی»[5] از دیدگاهی محاسباتی و با استفاده از مدل‌های آماری در پردازش زبان طبیعی بوده‌است. تعبیر «زبان سوگیرانه» یا «زبان مغرضانه» به استفاده از کلمات یا عبارت‌هایی اشاره دارد که به‌طور واضح یا ضمنی برای بیان جانبداری، تعصب، پیش‌داوری، تهاجم، اتهام، تحقیر (و مواردی از این دست) نسبت به موضوعات یا افراد به‌کارمی‌روند. این تعبیر در تقابل با «زبان غیرسوگیرانه»[6] یا «زبان بی‌طرفانه» قرارمی‌گیرد. با آنکه معیارهایی برای تشخیص استفاده از زبان سوگیرانه وجود دارد اما گاه تشخیص سوگیری زبانی حتی برای انسان نیز مشکل می‌گردد زیرا امکان دارد سوگیری دارای ساختاری زیرکانه و پیچیده باشد. این موضوع ریشه در پیچیدگی‌های معنایی خود زبان دارد و نویسنده/گوینده نیز می‌تواند با مهارت خود در بهره‌گیری از امکانات بالقوه زبان آن‌ها را در خدمت اهداف خود درآورده و برای تاثیرگذاری بیشتر بر مخاطب به‌کارگیرد یا فضا را برای قبولاندن مطالب و جهت‌دهی فکری به او هموار سازد.

۱-۲ ادبیات و پیشینه پژوهش

برای اجتناب از زبان سوگیرانه، معمولاً راهنماهای سبک نگارش فهرستی از کلمات را مشخص‌کرده و آن‌ها را به دلایلی نظیر ابهام، جانبداری از یک دیدگاه خاص یا چاپلوسانه‌بودن، سوگیرانه معرفی‌می‌نمایند و به نویسندگان توصیه‌می‌کنند که هنگام نگارش متون خود از آن‌ها

1 Natural language processing (NLP)
2 Classification
3 Clustering
4 Keywords extraction
5 Language bias
6 Unbiased Language

اجتناب نمایند. برای مثال دانش‌نامه ویکی‌پدیا -که هر کسی می‌تواند در آن بنویسد- نیز چنین راهنمایی را در اختیار کاربران خود قرارداده‌است و «سیاست دیدگاه خنثی»[1] را در نوشتن مقالات اِعمال‌می‌کند. به همین دلیل پیکره‌ای با نام «دیدگاه خنثی» یا «ان.پی.او.وی»[2] پدیدآمده‌است که شامل نسخه‌های مختلف دسته‌ای از مقالات ویکی‌پدیاست که از سوی کاربران سوگیرانه تشخیص‌داده‌شده و ویرایش شده‌اند. پس از تغییر یا حذف موارد سوگیرانه از متن مقالات ویکی‌پدیا، چندین ویرایش از آن مقاله پدیدمی‌آید، پیکره «ان.پی.او.وی» مجموعه این مقالات است که ویرایش‌های متعدد آن‌ها اختصاصاً به‌خاطر تغییر یا حذف کلمات و عبارات سوگیرانه از نسخه‌های قبلی به‌وجودآمده‌اند. این پیکره در مطالعه ریکازنس[3] و همکاران [۱] در استخراج «مدل‌های زبان‌شناسی برای تحلیل و شناخت سوگیری زبانی» مورد استفاده قرارگرفته‌است که هدف آن تشخیص خودکار سوگیری زبانی با استفاده از متن‌های سوگیرانه واقعی بوده‌است. بر مبنای نتایج این مطالعه می‌توان «انواع سوگیری‌های زبانی» را به دو گروه اصلی «سوگیری معرفت‌شناختی»[4] و «سوگیری قالبی»[5] تقسیم کرد. سوگیری معرفت‌شناختی شامل گزاره‌هایی است که معمولاً بر درست یا غلط بودن آن‌ها توافق وجوددارد و با زیرکی در متن برای ازپیش‌انگاری، استلزام یا تصریح یک موضوع به‌کارمی‌روند یا به‌عنوان تعدیل‌کننده و تخفیف‌دهندهٔ مسئولیت گوینده درباره یک مدعا، مورد استفاده قرارمی‌گیرند [۱]. سوگیری‌های معرفت‌شناختی را می‌توان به چند زیردسته مانند افعال واقعیت‌انگاری[6] [۲]، استلزام‌ها[7]، افعال تصریحی[8] [۳] و تردیدواژه‌ها[9] تقسیم کرد. ساختار سوگیری قالبی غالباً ساده‌تر از سوگیری معرفت‌شناختی بوده و به راحتی قابل تشخیص است زیرا زمانی اتفاق‌می‌افتد که نویسنده/گوینده از کلمات یا تعابیر سلیقه‌ای و یا از کلمات یک‌بعدی که فقط یک سوی یک واقعیت را نشان‌می‌دهند، استفاده‌نماید. چنین کلماتی گویای موضع شخص نویسنده/گوینده در قبال موضوع مورد بحث هستند [۴]. قابل ذکر است که می‌توان برای واژه‌هایی که دارای

1 Neutral Point Of View (NPOV) policy
2 NPOV Corpus
3 Recasens
4 Epistemological bias
5 Framing bias
6 Factive verbs
7 Entailments
8 Assertive verbs
9 Hedges

سوگیری قالبی هستند، قطبیت را نیز تعیین نمود. از انواع سوگیری‌های قالبی می‌توان به تشدیدکننده‌ها[1] و واژه‌های یک‌بعدی‌نگر[2] اشاره‌کرد. تشخیص سوگیری قالبی گاه با عنوان‌هایی نظیر «تشخیص موضع»[3] یا «تشخیص گرایش در بحث»[4] در مطالعات مطرح‌شده‌است [۵]، [۶] و [۷]، [۸] و [۹]. هدف چنین مطالعاتی تعیین و شناخت طرفی است که در یک مناقشه به نفع او موضع‌گیری و از دیدگاه او به موضوع نگاه‌شده‌است. برای نمونه در بعضی پژوهش‌ها یک دسته‌بندی دو حالته به صورت موضع مثبت (طرفدار شخص یا چیزی) یا موضع منفی (علیه شخص یا چیزی) مطرح‌شده و جملات یا اَسناد بر اساس آن تفکیک‌شده‌اند [۵]، [۶] و [۷]. گاهی نیز چند دیدگاه متضاد درنظرگرفته‌شده و جملات با توجه به نزدیکی یا جانبداری آن‌ها از هر یک از دیدگاه‌ها دسته‌بندی شده‌اند [۸] و [۹]. گاه ممکن است خواننده یک مقاله خبری از جوانب مختلف موضوع یا رویدادی که در خبر آورده‌شده‌است، آگاه نباشد. به همین دلیل پاتانکار[5] و بُس[6] در پژوهش خود [۱۰] سیستمی را آموزش داده و با استفاده از آن اقدام به کشف کلمات سوگیرانه دارای قطبیت مثبت/منفی در یک متن خبری کرده‌اند سپس بر اساس آن‌ها میزان سوگیری زبانی خبر را محاسبه نموده‌اند. این سیستم با استفاده از شیوه‌های پردازش زبان طبیعی در استخراج کلمات کلیدی متن، موضوع خبر را تعیین‌کرده و آن‌گاه سعی‌می‌کند خبرهایی با همان موضوع اما با سوگیری‌های متفاوت را در وب بیابد و به خواننده خبر پیشنهاد دهد.

طبق اطلاعات نگارندگان تاکنون پژوهش زبان‌شناسی با رویکرد رایانشی به منظور تحلیل و بررسی سوگیری رسانه‌ای در خبرهای فارسی انجام‌نشده و پیکره(های) لازم برای به‌کارگیری روش‌های بانظارت در انجام چنین تحقیقاتی نیز برای زبان فارسی موجود نمی‌باشد بنابراین در این پژوهش سعی بر آن بوده‌است که با استفاده از مدل‌های آماری به تحلیلی برای «سوگیری زبانی» رسانه‌ها دست‌یافته و پیکره‌ای از «واژگان سوگیرانه فارسی»[7] نیز در این راستا تهیه شود تا در آینده برای انجام چنین پژوهش‌هایی در دسترس سایر محققان قرار گیرد. در تحقیق

1 Intensifiers
2 One-sided terms
3 Stance recognition
4 Recognizing Arguing Subjectivity
5 Patankar
6 Bose
7 Persian bias lexicon

حاضر بررسی سوگیری زبانی از طریق کنکاش در مجموعه‌ای از متون خبری فارسی انجام‌شده‌است که از وب‌گاه چهار بنگاه خبری بین‌المللی دویچه‌وله[1]، رادیو بین‌المللی فرانسه[2]، اسپوتنیک[3] و العربیه[4] در دو بازه زمانی یکسان -حدوداً دو هفته‌ای- قبل و بعد از تصویب «برجام: برنامه جامع اقدام مشترک»[5]-، گردآوری شده‌اند. به زبان بسیار ساده این پژوهش به این سوال پاسخ داده‌است که هر یک از بنگاه‌های خبری مورد بررسی در بازه‌های زمانی تعیین‌شده موضوعات خبریِ منتخب خود را به چه زبانی بیان‌کرده‌اند (استفاده از زبان سوگیرانه یا غیرسوگیرانه).

۲ گردآوری داده و آماده‌کردن پیکره

برای انجام این پژوهش در ابتدا نیاز به یک مجموعه داده و به عبارت بهتر پیکره‌های خبری داریم که بر اساس اخبار فارسی منتشرشده در بنگاه‌های خبری جمع‌آوری و مرتب‌سازی شده‌باشند، به این منظور چهار بنگاه خبری دویچه‌وله، رادیو بین‌المللی فرانسه (آراف‌آی)، اسپوتنیک و العربیه که زبان اصلی پخش خبر آن‌ها به ترتیب آلمانی، فرانسوی، روسی و عربی می‌باشد، انتخاب شده و مجموعه‌ای از متون خبری مرتبط با ایران از وبگاه فارسی آن‌ها جمع‌آوری گردیده‌است. حجم پیکره جمع‌آوری‌شده از لحاظ تعداد خبرها و کلمات، به تفکیک بنگاه خبری و بازه زمانی در جدول ۲ آمده‌است.

جدول ۲: حجم پیکره جمع‌آوری‌شده برای بررسی سوگیری رسانه‌ای

بنگاه خبری	بازه زمانی	تعداد اخبار	تعداد کلمات
آراف‌آی	پیش از برجام	۸۴	۳۱۱۶۲
	پس از برجام	۵۶	۱۹۱۱۵
اسپوتنیک	پیش از برجام	۹۲	۳۰۴۴۲
	پس از برجام	۸۶	۳۳۹۱۱

1 Deutsche Welle (DW)
2 Radio France Internationale (RFI)
3 SPUTNIK
4 Al Arabiya
5 JCPOA: Joint Comprehensive Plan of Action

بنگاه خبری	بازه زمانی	تعداد اخبار	تعداد کلمات
العربیه	پیش از برجام	۱۲۴	۴۵۵۵۶
	پس از برجام	۶۳	۲۲۷۲۶
دویچه‌وله	پیش از برجام	۱۶۲	۱۱۴۸۶۹
	پس از برجام	۱۱۷	۷۸۶۹۸
جمع کل		۷۸۴	۳۷۶۴۷۹

برای تهیه هدفمند خبرها، تاریخ دستیابی به «برنامه جامع اقدام مشترک» (برجام) که برابر با ۱۴ جولای ۲۰۱۵ بوده‌است به‌عنوان یک نقطه عطف و نقطه میانی یک دوره ۴ماهه انتخاب‌گردیده و اخبار نوشتاری همه خبرگزاری‌های مذکور در دو بازه زمانی حدوداً دو هفته‌ای که از ۱/۵ ماه قبل و ۱/۵ ماه بعد از تاریخ برجام شروع می‌شود (یعنی چهاردهم تا سی‌ویکم ماه می‌و یکم تا چهاردهم ماه سپتامبر سال ۲۰۱۵)، برداشته‌شده‌اند. هدف از انتخاب این دو تاریخ، بررسی تغییر احتمالی سیاست‌ها و رویکردهای این بنگاه‌های خبری نسبت به ایران با استفاده از مقایسه میزان سوگیری‌های احتمالی آن‌ها، قبل و پس از امضای برجام بوده‌است که در زبان به‌کاربرده‌شده در متون خبری نمود پیداکرده‌است. همچنین با جستجوی کلیدواژه «ایران» در آرشیو اخبار نوشتاری بنگاه‌های خبری مذکور، فقط خبرهایی که در متن یا عنوان آن‌ها از این کلمه استفاده شده‌بوده‌است، گردآوری شده‌اند. دلیل این کار، قبول این فرضیه از سوی پژوهشگر است که با تکیه بر حضور واژه «ایران» در خبر می‌توان نتیجه‌گرفت که موضوع یا رویداد مطرح‌شده در خبر با کشور ایران مرتبط می‌باشد.

از جمله چالش‌های جمع‌آوری داده می‌توان به طولانی و زمان‌بر بودن جمع‌آوری داده به صورت دستی و عدم امکان استفاده از خزشگرها به‌خاطر ندادن اجازه دسترسی از طرف بعضی از وب‌سایت‌ها اشاره‌کرد.

۲-۱ انجام مراحل پیش‌پردازشی

به‌منظور آماده‌سازی داده‌ها برای پردازش لازم است که همه آن‌ها را به قالبی یکسان و قابل پذیرش برای نرم‌افزار انتخابی درآوریم تا برای پردازش‌های بعدی آماده گردند. برای

دستیابی به این هدف ابتدا پیش‌پردازش‌هایی بر روی متون انجام‌گرفته است که از آن جمله می‌توان به یکسان‌سازی بعضی از اختلافات در رسم‌الخط کلمات، مرتب‌سازی متن‌ها و از بین‌بردن فواصل اضافی که باعث ناهمگونی متون می‌گردند و حذف علامت‌های نگارشی اشاره‌کرد. پس از انجام این مراحل نیز نرمال‌سازی متن‌ها با استفاده از ابزار نرمال‌ساز شرکت عصرگویش‌پرداز انجام‌شده‌است [۱۱]. همچنین با استفاده از یک برچسب‌زن مقوله‌های نحوی که دارای دقت میانگین ٪۹۴ است [۱۲]، برچسب مقوله نحوی نیز به کلمات پیکره‌های خبری تخصیص‌داده‌شده‌است تا از این ویژگی در مدل احتمالاتی برای کمک به تشخیص کلمات سوگیرانه در متن استفاده‌گردد.

مرحله پیش‌پردازش داده‌ها از اهمیت خاصی برخوردار است زیرا مستقیماً در درستی و دقت نتایج تاثیرگذار می‌باشد. از جمله چالش‌هایی که به هنگام پیش‌پردازش داده‌ها با آن روبرو می‌شویم می‌توان به وجود اشتباهات املایی در متن خبرها، یکسان نبودن رسم‌الخط متن‌های خبری و همچنین پیداکردن برنامه‌هایی با دقت قابل قبول برای استخراج ستاک[1] و لِم[2] کلمات اشاره کرد.

۲-۲ آماده‌کردن پیکره واژگان سوگیرانه فارسی

برای شناسایی واژه‌هایی که مصداق زبان سوگیرانه محسوب‌می‌شوند نیاز به فهرستی از واژه‌های سوگیرانه داریم تا بتوانیم بر مبنای آن‌ها تصمیم‌گیری نماییم اما (طبق دانسته‌های نگارندگان) متاسفانه تا قبل از این پژوهش چنین واژگانی برای زبان فارسی موجود نبوده‌است. به همین دلیل یکی از کارهایی که در خلال این پژوهش انجام‌شده تهیه پیکره‌ای از واژه‌های سوگیرانه فارسی بوده‌است. ویژگی مشترک همگی واژه‌هایی که در این پیکره گنجانده شده‌اند آن است که حداقل در یکی از معانی خود سوگیرانه تلقی‌می‌شوند. واژگان این فهرست با درنظرگرفتن معیارهایی که در پژوهش ریکازِنس و همکاران [۱] برای شناخت «انواع سوگیری زبانی» مطرح شده‌اند (و در این نوشتار نیز به آن‌ها اشاره‌شد) و بر مبنای مجموعه واژگان سوگیرانه‌ای که در آن پژوهش [۱] برای زبان انگلیسی استخراج گشته، با اعمال تغییرات لازم

1 Stem
2 Lemma

برای کاربردی کردن آن‌ها برای زبان فارسی و مناسب‌سازی آن‌ها برای متـون خبـری، انتخـاب شده‌اند. «پیکره واژگان سوگیرانه فارسی» که در نهایت آماده‌شده‌است شامل حـدود ۵۸۰۰ واژه می‌باشد.

۳-۲ برچسب‌دهی بخشی از کلمات پیکره برای آموزش مدل

بخش‌هایی از متن هر کدام از هشت پیکره خبری - متشکل از اخبار چهار بنگاه خبری در دو بازه زمانی قبل و پس از برجام- انتخاب شده و به صورت دستی توسط عامل انسانی خبره با دو عنوان «سوگیرانه» یا «غیرسوگیرانه» برچسب‌گذاری شده‌اند. این مجموعه با حدود ۴۶۰۰۰ کلمهٔ برچسب‌خورده -که تقریبا دوازده درصد از کل کلمات پیکره را شامل می‌شود- خود یک پیکره جداگانه را تشکیل داده و برای ساختن مدلی احتمالاتی به شیوه بانظارت جهت پیش‌بینی کلمات سوگیرانه متن‌ها، مورد استفاده قرارگرفته‌است. در جدول ۳ نمونه‌ای از پیکرهٔ متن‌های خبری برچسب‌خورده -که از خبرهای بنگاه اسپوتنیک در بازه زمانی قبل از برجام برداشته شده- آورده‌شده‌است. در این پیکره، کلمات «سوگیرانه» با برچسب ’b‘ و کلمات «غیرسوگیرانه» با برچسب ’n‘ مشخص‌شده‌اند.

جدول ۳: نمونه‌ای از پیکره برچسب‌خورده با دو عنوان ’b‘ و ’n‘

واحد متنی	امکان	رسیدن	به	توافق	با	گروه
برچسب	b	b	n	b	n	n
واحد متنی	۵	+	۱	بسیار	زیاد	است
برچسب	n	n	n	b	b	n

۳ استخراج ویژگی‌های زبانی از پیکره

برای به دست آوردن میزان سوگیری زبانی در پیکره‌های خبری، بر اساس تعریف چندین ویژگیِ زبانی و متنی و بررسی اینکه آیا هر کلمه از پیکره دارای ویژگی‌های مورد نظر هسـت یـا خیر، عمل شده‌اسـت. بنـابراین بـرای هـر یـک از کلمـات، بـرداری متشـکل از هشـت ویژگـی مقداردهی شده و سپس با کنار هم قراردادن بردارهای ویژگی کلمات هر پیکره، ماتریس مربوط به آن پیکره خبری استخراج گشته است. هشت ویژگی که پس از بررسـی آن‌هـا مقـادیر بـردار

ویژگی کلمه تخصیص می‌یابند، به شرح زیر می‌باشند:

۱- آیا کلمه در پیکره واژگان سوگیرانه فارسی موجود است؟ {صفر/یک}.

اولین ویژگی‌ای که برای هر کدام از واژه‌های متون خبری بررسی می‌شود و در بردار ویژگی‌ها تخصیص می‌یابد، بودن یا نبودن واژه مورد نظر در «پیکره واژگان سوگیرانه فارسی» است.

۲- آیا کلمه در پیکره واژگان دارای برچسب قطبیت مثبت موجود است؟ {صفر/یک}.

۳- آیا کلمه در پیکره واژگان دارای برچسب قطبیت منفی موجود است؟ {صفر/یک}

از لحاظ نظری، واژگانی که می‌توان برچسب مثبت/منفی (غیرخنثی) از نظر قطبیت به آن‌ها تخصیص‌داد، زیرمجموعه‌ای از واژگان سوگیرانه هستند و می‌توان آن‌ها را در زمره واژه‌های دارای سوگیری قالبی محسوب‌نمود. به کمک پیکره «واژگان فارسی دارای برچسب قطبیت» [۱۳] -که شامل تعدادی از صفات است که بر مبنای قطبیت مثبت/منفی برچسب‌دهی شده‌اند- ویژگی دیگری در بردار ویژگی‌ها برای کلمات پیکره مقداردهی می‌شود که حضور یا غیبت آن کلمه در میان واژگان دارای قطبیت مثبت یا منفی است.

۴- آیا کلمه از مقوله نحوی صفت یا قید هست؟ {صفر/یک}. یک دسته از واژه‌های سوگیرانه قالبی «تشدیدکننده‌ها» هستند که برای ایجاد یا تشدید تمایل/ تنفر مخاطب نسبت به موضوعی، در کلام آورده‌می‌شوند و چون دارای قطبیت مثبت/منفی هستند اگر در خبری به‌کاربرده شوند، نشان‌دهنده موضع نویسنده خبر می‌باشند. از آنجا که این تشدیدکننده‌ها از لحاظ نحوی اغلب در زمره صفت‌ها و قیدها قرارمی‌گیرند بنابراین یکی دیگر از ویژگی‌هایی که در بردار ویژگی کلمات به آن توجه شده‌است، مقوله نحوی کلمه است. به این ترتیب در هنگام تخصیص مقادیر بردار ویژگی هر کلمه، اگر کلمه دارای برچسب 'ADJ'، 'ADV' یا 'QUA'[1] باشد ویژگی مرتبط با مقوله نحوی کلمه با مقدار یک و در غیر این صورت با صفر مقداردهی می‌شود.

۵- آیا هیچ‌یک از کلماتی که در متن در فواصل ۱± یا ۲± از کلمه هدف قرار گرفته‌اند، در پیکره واژگان سوگیرانه فارسی موجودند؟ {صفر/یک}.

ویژگی دیگری که در بردار ویژگی کلمات در ارتباط با پیکره واژگان سوگیرانه

---

1 برچسب‌های مربوط به صفت‌ها، قیدها و کیفیت‌نماها در پیکره بی‌جن‌خان

تعریف‌شده‌است بررسی وجود یک کلمه سوگیرانه در فاصله یک یا دوتایی کلمه اصلی است زیرا آن‌طور که در نتایج مطالعه ریکازِنس و همکاران [۱] دیده‌شده‌است کلمات سوگیرانه اغلب در خوشه‌هایی، در اطراف یکدیگر رخ می‌دهند.

۶- آیا هیچ‌یک از کلماتی که در متن در فواصل ۱± یا ۲± از کلمه هدف قرار گرفته‌اند، در پیکره واژگان دارای برچسب قطبیت مثبت موجودند؟ {صفر/یک}

۷- آیا هیچ‌یک از کلماتی که در متن در فواصل ۱± یا ۲± از کلمه هدف قرار گرفته‌اند، در پیکره واژگان دارای برچسب قطبیت منفی موجودند؟ {صفر/یک}

دو ویژگی دیگری که در ارتباط با قطبیت کلمات تعریف می‌شوند و مربوط به بافت متنی اطراف کلمه می‌باشند، بررسی حضور/عدم حضور یک واژه با قطبیت مثبت/منفی (غیرخنثی) در فاصله یک یا دوتایی اطراف کلمه هدف است. این دو ویژگی و همچنین ویژگی دیگر وابسته به بافت که در شماره ۵ گفته‌شد از جمله خصوصیاتی هستند که به باهم‌آیی‌های کلمات مربوط می‌شوند و تلاشی در جهت افزایش (احتمالی) مقدار معیار «بازخوانی»[1] مدل می‌باشند زیرا باعث می‌شوند احتمالی برای سوگیرانه بودن کلماتی که در هیچ‌یک از پیکره‌های سوگیرانه یا قطبیت دیده‌نشده‌اند، در مدل لحاظ شود.

۸- برچسب کلمه از لحاظ سوگیرانه یا غیرسوگیرانه بودن کدام است؟ {b (سوگیرانه)، n (غیرسوگیرانه)، ? (نامشخص)}

برچسب کلمه در واقع نمایانگر دسته‌ای است که کلمه به آن تعلق دارد. این ویژگی می‌تواند دارای مقادیر مشخص 'b' (سوگیرانه) و 'n' (غیرسوگیرانه) برای کلمات برچسب‌خورده یا مقدار '?' (نامشخص) برای کلمات بدون برچسب باشد. این ویژگی در فرایند یادگیری بانظارت به عنوان کلاس یا دستهٔ هدف مورد استفاده قرار می‌گیرد.

۴ نتایج به‌دست‌آمده

با استفاده از هشت ویژگی مذکور، ماتریسی از بردارهای ویژگی پیکره کلمات برچسب‌خورده تشکیل‌شده‌است. سپس این ماتریس به ابزار وکا[2] [۱۴] داده‌شده تا به کمک یک

---

1 Recall
2 WEKA

طبقه‌بندی‌کننده، مدلی برای دسته‌بندی کلمات به سوگیرانه و غیرسوگیرانه بر مبنای آن استخراج گردد. مدل مذکور با استفاده از تکنیک «اعتبارسنجی متقاطع ده لایه»[1] بر روی مجموعه داده‌های برچسب‌خورده، آموزش داده‌شده و سپس مورد آزمون قرارگرفته‌است.

برای به‌دست‌آوردن مدل احتمالاتی مناسب، چندین دسته‌بندی‌کننده از ابزار وکا بر روی داده‌های دارای برچسب سوگیرانه/غیرسوگیرانه به شیوه اعتبارسنجی متقاطع ده لایه مورد آزمایش قرارگرفته‌اند. از آنجا که معیارهای دقت[2] و بازخوانی معمولا در مصالحه با یکدیگر هستند بنابراین باید بین آن‌ها به نوعی تعادل برقرارنمود. به همین دلیل پس از مقایسه مقادیر ارزیابی «معیار اف»[3] و «سطح زیر منحنی راک»[4] برای سه طبقه‌بندی‌کننده «بیزی»[5]، «رگرسیون لجستیک»[6] و «ماشین بردار پشتیبان»[7]، با توجه به بهتربودن نتایج ارزیابی دسته‌بندی‌کننده «بیزی»، این طبقه‌بندی‌کننده برای آموزش مدل برگزیده شده‌است. همان‌طورکه در جدول ۴ دیده‌می‌شود مقادیر میانگین کل برای معیارهای ارزیابی دقت، بازخوانی و معیار اِف در این مدل در هنگام دسته‌بندی کلمات به سوگیرانه یا غیرسوگیرانه به ترتیب حدود ٪۷۶، ٪۷۷ و ٪۷۶ می‌باشد.

جدول ۴: مقادیر معیارهای ارزیابی برای دسته‌بندی‌کننده بیزی

برچسب دسته	سطح زیر منحنی راک	معیار اف	بازخوانی	دقت
غیرسوگیرانه	۰,۷۴	۰,۸۲۳	۰,۸۷۳	۰,۷۷۹
سوگیرانه	۰,۷۴	۰,۶۵۷	۰,۵۹۲	۰,۷۳۹
میانگین وزن‌دار	۰,۷۴	۰,۷۶۱	۰,۷۶۷	۰,۷۶۴

بر طبق ارزیابی نرم‌افزار وکا، رتبه‌بندی مشخصه‌هایی که برای تشکیل ماتریس بردارهای ویژگی کلمات مورد استفاده قرارگرفته‌اند، به صورتی است که در جدول ۵ آمده‌است:

1 10-fold cross-validation
2 Precision
3 F-Measure
4 Area under the ROC curve
5 Bayesian classifier
6 Logistic regression
7 Support Vector Machine

جدول ۵: رتبه‌بندی ویژگی‌های مورد استفاده در مدل بر اساس ارزیابی نرم‌افزار

رتبه	ویژگی
۱	حضور کلمه در پیکره واژگان سوگیرانه فارسی
۲	مقوله نحوی کلمه (صفت یا قید)
۳	حضور کلمه در پیکره واژگان دارای برچسب قطبیت منفی
۴	حضور کلمه در پیکره واژگان دارای برچسب قطبیت مثبت
۵	حضور کلمه‌هایی که در پیکره واژگان سوگیرانه فارسی هستند در اطراف کلمه هدف (فاصله ۱± یا ۲±)
۶	حضور کلمه‌هایی که در پیکره واژگان دارای برچسب قطبیت مثبت هستند در اطراف کلمه هدف (فاصله ۱± یا ۲±)
۷	حضور کلمه‌هایی که در پیکره واژگان دارای برچسب قطبیت منفی هستند در اطراف کلمه هدف (فاصله ۱± یا ۲±)

ترتیب موجود در این رده‌بندی بدان معناست که هر کدام از ویژگی‌های درنظرگرفته‌شده تا چه اندازه قدرت پیش‌بینی برای مدل ایجادکرده‌اند.

۴-۱ پیش‌بینی کلمات سوگیرانه در متن‌های خبری بدون برچسب

پس از حذف علامت‌های نگارشی و اعداد از مجموعه داده‌ها، در این مرحله با استفاده از بردار ویژگی‌های کلمات بدون برچسب پیکره‌های خبری، هشت ماتریس (به ازای چهار بنگاه خبری در دو بازه زمانی قبل و پس از برجام) ساخته‌شده‌است. تفاوت این ماتریس‌ها با ماتریس بردارهای ویژگی کلماتی که دارای برچسب سوگیرانه/غیرسوگیرانه هستند در نحوه مقداردهی به ویژگی آخر است زیرا به جای تخصیص یکی از مقادیر ’b‘ یا ’n‘ که نشانگر تعلق کلمه به دسته کلمات سوگیرانه یا غیرسوگیرانه است، از کاراکتر ’?‘ به معنای مقدار «نامشخص» یا «تخصیص‌نیافته» استفاده‌می‌شود. در آخر، هر یک از این هشت ماتریس به‌صورت جداگانه به ابزار وکا واردشده‌اند تا به کمک مدلی که در مرحله قبل ساخته‌شده‌است برچسب ’b‘ یا ’n‘ برای کلمات بدون برچسب پیکره‌های خبری پیش‌بینی شود.

۴-۲ محاسبه نرخ سوگیری زبانی در پیکره‌های خبری

پس از تعیین کلمات سوگیرانه متن‌ها به کمک مدل احتمالاتی، درصد استفادهٔ هر بنگاه خبری از واژه‌های سوگیرانه در بازه زمانی مشخص‌شده قابل تخمین است. به این ترتیب با تقسیم تعداد کلمات سوگیرانه هر پیکره خبری به کل کلمات آن، نرخ سوگیری زبانی برای هر پیکره محاسبه‌گردیده‌است. مقادیر نرخ سوگیری زبانی به تفکیک بازه زمانی در جدول ۶ و جدول ۷ قابل مشاهده‌اند.

جدول ۶: نرخ سوگیری زبانی در بازه زمانی پیش از برجام (مربوط به پیکره‌های خبری برچسب‌خورده به کمک مدل احتمالاتی)

بنگاه خبری	تعداد کلمات سوگیرانه پیکره	تعداد کل کلمات پیکره	نرخ سوگیری زبانی
آراف‌آی	۶۹۴۱	۲۲۷۹۰	۳۰٪
اسپوتنیک	۷۲۰۹	۲۲۳۴۳	۳۲٪
العربیه	۱۰۲۲۷	۳۵۵۱۵	۲۹٪
دویچه‌وله	۲۸۶۶۶	۹۶۸۳۲	۳۰٪

جدول ۷: نرخ سوگیری زبانی در بازه زمانی پس از برجام (مربوط به پیکره‌های خبری برچسب‌خورده به کمک مدل احتمالاتی)

بنگاه خبری	تعداد کلمات سوگیرانه پیکره	تعداد کل کلمات پیکره	نرخ سوگیری زبانی
آراف‌آی	۳۸۲۵	۱۲۲۱۳	۳۱٪
اسپوتنیک	۷۳۱۱	۲۲۸۶۲	۳۲٪
العربیه	۴۷۰۵	۱۵۴۹۴	۳۰٪
دویچه‌وله	۲۰۰۸۹	۶۴۸۳۹	۳۱٪

همچنین نرخ‌های سوگیری زبانی محاسبه‌شده برای قسمت‌هایی از پیکره‌های خبری که با قضاوت انسانی برچسب‌دهی شده‌اند، در جدول ۸ و جدول ۹ قابل مشاهده‌اند.

جدول ۸: نرخ سوگیری زبانی در بازه زمانی پیش از برجام (مربوط به پیکره‌های خبری برچسب‌خورده با قضاوت انسانی)

بنگاه خبری	تعداد کلمات سوگیرانه پیکره	تعداد کل کلمات پیکره	نرخ سوگیری زبانی
آراف‌آی	۱۶۵۷	۴۴۹۲	۳۷٪
اسپوتنیک	۱۱۷۵	۵۰۷۸	۲۳٪
العربیه	۲۱۱۷	۴۹۴۰	۴۳٪
دویچه‌وله	۱۹۹۲	۴۵۵۷	۴۴٪

جدول ۹: نرخ سوگیری زبانی در بازه زمانی پس از برجام (مربوط به پیکره‌های خبری برچسب‌خورده با قضاوت انسانی)

بنگاه خبری	تعداد کلمات سوگیرانه پیکره	تعداد کل کلمات پیکره	نرخ سوگیری زبانی
آراف‌آی	۲۰۲۹	۴۷۱۰	۴۳٪
اسپوتنیک	۲۴۲۸	۷۹۸۲	۳۰٪
العربیه	۲۰۶۵	۴۵۹۸	۴۵٪
دویچه‌وله	۲۰۴۸	۴۷۲۹	۴۳٪

۴-۳مقایسه نرخ سوگیری زبانی بنگاه‌های خبری در دو بازه زمانی قبل و پس از برجام

در آخر، نرخ سوگیری زبانی بنگاه‌های خبری با استفاده از میانگین‌گیری وزن‌دار بین نرخ سوگیری متون خبری با برچسب انسانی و متون خبری با برچسب ماشینی (احتمالاتی) محاسبه‌شده و مقادیر به‌دست‌آمده در جدول ۱۰ آورده شده‌اند.

جدول ۱۰: نرخ سوگیری زبانی هر یک از بنگاه‌های خبری با استفاده از میانگین‌گیری وزن‌دار

بنگاه خبری	بازه زمانی	
	پیش از برجام	پس از برجام
آراف‌آی	۳۲٪	۳۵٪
اسپوتنیک	۳۱٪	۳۲٪
العربیه	۳۱٪	۳۴٪
دویچه‌وله	۳۰٪	۳۲٪

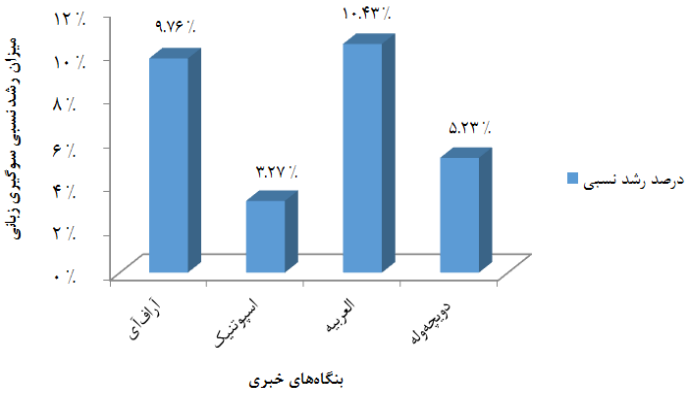
با مقایسه نتایج به‌دست‌آمده مشخص می‌گردد که در هر چهار بنگاه خبری میزان استفاده از کلمات سوگیرانه در خبرها در بازه زمانی پس از برجام نسبت به بازه زمانی پیش از آن رشد داشته‌است.

در همین راستا درصد رشد نسبی سوگیری زبانی هر بنگاه خبری در بازه زمانی پس از برجام نسبت به بازه زمانی پیش از آن به روش زیر محاسبه‌شده‌است:

$$R = \frac{\text{نرخ سوگیری زبانی بنگاه خبری در بازه زمانی پس از برجام}}{\text{نرخ سوگیری زبانی بنگاه خبری در بازه زمانی پیش از برجام}} \quad (۱)$$

$$\text{درصد رشد نسبی سوگیری زبانی} = (۱ - R) * ۱۰۰ \quad (۲)$$

مقدار درصد رشد نسبی سوگیری زبانی هر یک از بنگاه‌های خبری بر روی نمودار شکل ۲ نشان‌داده شده است.



شکل ۲: نمودار درصد رشد نسبی سوگیری زبانی هر بنگاه خبری در بازه زمانی پس از برجام نسبت به بازه زمانی پیش از آن

با مراجعه به شکل ۲ مشاهده‌می‌شود که ترتیب به‌دست‌آمده برای رده‌بندی بنگاه‌های خبری از لحاظ میزان رشد نسبی سوگیری زبانی به‌صورت: العربیه، آراف‌آی، دویچه‌وله و اسپوتنیک می‌باشد. به بیان دیگر اگر میزان رشد نسبی سوگیری زبانی اسپوتنیک را که کمترین مقدار است به عنوان مبنا درنظربگیریم می‌توان گفت که میزان رشد نسبی سوگیری زبانی بنگاه خبری دویچه‌وله ۱/۶ برابر، آراف‌آی ۳ برابر و العربیه ۳/۲ برابر اسپوتنیک بوده‌است.

۵ جمع‌بندی و نتیجه‌گیری

در این پژوهش رخداد یکی از انواع سوگیری رسانه‌ای به نام «سوگیری زبانی» در متون خبری فارسی چهار بنگاه خبری بین‌المللی آراف‌آی، اسپوتنیک، العربیه و دویچه‌وله در دو بازه زمانی -تقریبا دو هفته‌ای- قبل و بعد از تصویب برجام مورد بررسی قرارگرفت. در این راستا با به‌کارگیری الگوریتم‌های یادگیری ماشینی بانظارت به پردازش و تحلیل متن‌های خبری برای شناسایی کلمات سوگیرانه متن و محاسبه نرخ سوگیری اقدام‌گردید. پس از مقایسه نرخ سوگیری زبانی به‌دست‌آمده برای هر یک از بنگاه‌های خبری مورد مطالعه در دو بازه زمانی قبل و پس از برجام مشاهده‌شد که این نرخ در همه بنگاه‌های خبری در دوره زمانی پس از برجام نسبت به دوره زمانی قبل از آن افزایش داشته‌است و اگر بخواهیم بنگاه‌های خبری را بر اساس میزان افزایش نرخ سوگیری زبانی از بزرگ به کوچک مرتب کنیم به‌صورت العربیه (٪۱۰,۴۳)، آراف‌آی (٪۹,۷۶)، دویچه‌وله (٪۵,۲۳) و اسپوتنیک (٪۳,۲۷) می‌باشند.

یکی از کاربردهای این پژوهش، بررسی تغییر احتمالی رویکرد این بنگاه‌های خبری نسبت به ایران قبل و پس از امضای برجام می‌باشد. دلیل این موضوع آن است که بنگاه‌های خبری می‌توانند برای بیان یک خبر یکسان کلمات متفاوتی را به‌کاربرند که نشان‌دهنده گرایش و جهت‌گیری آن‌هاست. برای نمونه آن‌ها می‌توانند از الفاظ سوگیرانه «خرابکار» و «کشته» به جای کلمات غیرسوگیرانه «معترض» و «مرده» استفاده نمایند. بنابراین می‌توان نتیجه‌گرفت که هرچه میزان استفاده از کلمات سوگیرانه در بیان مطلبی کمتر باشد، امکان این‌که گوینده مطلب در موضع بی‌طرفی قرارداشته‌باشد بیشتر است.

برای انجام این پژوهش با محدودیت‌هایی نیز مواجه بودیم که از جملهٔ آن‌ها نبود پیکره‌ای از واژگان سوگیرانه فارسی برای آموزش مدل بود که به ایجاد آن اقدام گردید. محدودیت دیگر،

دسترسی نداشتن به ابزاری جهت استخراج لم کلمات با دقت قابل قبول بود که باعث شد از شناسایی بسیاری از افعال، اسامی یا صفات سوگیرانه بازمانیم زیرا صورت‌های تصریفی این کلمات بیش از آن هستند که بتوان همگی آن‌ها را در پیکره واژگان سوگیرانه جای‌داد. همین موضوع یکی از دلایلی بود که باعث شد معیار بازخوانی مدل از حدود ٪۶۰ فراتر نرود. در صورت برطرف شدن این مشکل علاوه بر امکان افزایش بازخوانی مدل، پیکره واژگان سوگیرانه نیز کوچک‌تر می‌گردد زیرا به جای صورت‌کلمه‌ها، لم کلمات را در خود جای خواهدداد.

منابع

[26] Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013, August). Linguistic Models for Analyzing and Detecting Biased Language. In *ACL (1)*(pp. 1650-1659).

[27] Kiparsky, P., & Kiparsky, C. (1968). *Fact*. Linguistics Club, Indiana University.

[28] Hooper, J. B. (1974). *On assertive predicates*. Linguistics Club, Indiana University.

[29] Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of communication, 57(1), 163-173.*

[30] Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., & Minor, M. (2011, June). Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis* (pp. 1-9). Association for Computational Linguistics.

[31] Conrad, A., & Wiebe, J. (2012, July). Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics* (pp. 80-88). Association for Computational Linguistics.

[32] Somasundaran, S., & Wiebe, J. (2010, June). Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 116-124). Association for Computational Linguistics.

[33] Yano, T., Resnik, P., & Smith, N. A. (2010, June). Shedding (a

thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 152-158). Association for Computational Linguistics.

[34] Park, S., Lee, K., & Song, J. (2011, June). Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 340-349). Association for Computational Linguistics.

[35] Patankar, A. A., & Bose, J. (2016, June). Bias Based Navigation for News Articles and Media. In *International* Conference *on Applications of Natural Language to Information Systems* (pp. 465-470). Springer International Publishing.

[۳۶] بحرانی، محمد، صامتی، حسین، حافظی، نازیلا، ممتازی، سعیده و موثق، حامد، (۱۳۸۵). "به‌کارگیری پیکره متنی زبان فارسی در ساخت مدل‌های زبانی آماری برای سیستم‌های بازشناسی گفتار پیوسته فارسی"، *دومین کارگاه پژوهشی زبان فارسی و رایانه*، صص ۱۰۹-۹۲.

[37] https://github.com/shohre10539/persian_pos_tagger

[38] Dehdarbehbahani, I., Shakery, A., & Faili, H. (2014). Semi-supervised word polarity identification in resource-lean languages. *Neural Networks*, *58*, 50-59.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

ارائه مدلی جهت خطایابی نحوی هوشمند در زبان فارسی با استفاده از دستورزبان وابستگی

پژمان مختاری فرد جونقانی* ، محمداحسان بصیری و ایمان مختاری فرد*****

چکیده

در دنیای امروز با افزایش زندگی ماشینی و ورود کامپیوتر در زندگی بشر و همچنین افزایش سندهای متنی الکترونیکی، ایجاد یک خطایاب جهت بررسی متون در زبان فارسی، اهمیت ویژه‌ای پیدا کرده است. یکی از انواع خطایابی موجود در متون، خطایابی نحوی است. خطایابی نحوی متون به منظور رسیدن به یک متن عاری از خطای نحوی است تا درک زبان برای ماشین افزایش یابد. در این مقاله با کمک از دستورزبان وابستگی به دنبال ارائه مدلی برای خطایابی نحوی متون هستیم. دستور زبان وابستگی، یکی از روش‌های نمایش دستور زبان در پردازش زبان طبیعی است که در این مقاله از آن استفاده می‌شود. مدل پیشنهادی این مقاله ابتدا اسناد دریافتی را نرمال و استاندارد می‌کند. سپس به جداسازی جملات و کلمات در متون می‌پردازد. در مرحله بعد، چهار ویژگی برچسب مقوله‌ای، تعداد، شخص و زمان را در کلمات مشخص می‌کند. سپس با استفاده از دستور زبان وابستگی جملات را بر اساس پنج شاخص نحوی تعریف شده، مورد بررسی و تحلیل قرار می‌دهد و در نهایت در صورت تشخیص خطا، آن‌ها را اعلام می‌کند. در ارزیابی این مدل برای تشخیص خطای نحوی در دو مجموعه جملات آزمایشی که شامل ۴۰۵ جمله نمونه از قبیل جملات بدون خطا و جملات با خطا بودند، میزان دقت[1] ۹۱٪ به دست آمد. استفاده از این مدل باعث کاهش ابهام در تولید متون در ترجمه ماشینی و افزایش درک زبان توسط ماشین می‌گردد.

واژه‌های کلیدی: پردازش زبان طبیعی، دستور زبان وابستگی، پردازش نحوی، خطایابی نحوی

---

* دانشجوی کارشناسی ارشد، موسسه آموزش عالی علوم و فناوری سپاهان، Pejman@snu.ir

** هیأت علمی مهندسی نرم افزار دانشگاه شهرکرد، Basiri.sku@gmail.com

*** دانشجوی دکتری مهندسی نرم افزار دانشگاه آزاد اسلامی واحد قم

1 Accuracy

۱- مقدمه

امروزه ماشین‌ها هر چه بیشتر زبان انسان را متوجه شوند، کار کردن با آن‌ها راحت‌تر است و کارایی و سرعت استفاده از سرویس‌های آن‌ها بیشتر است. با توجه به این مطلب، پردازش زبان طبیعی[1] که یکی از شاخه‌های هوش مصنوعی است، به‌عنوان یک نیاز و رویکرد مثبت برای ارتباط بین انسان و ماشین شناخته می‌شود. زبانی که انسان برای برقراری ارتباط از آن استفاده می‌کند را زبان طبیعی می‌گویند [۱].

برای ایجاد این ارتباط، به دو پردازش مهم در کامپیوتر نیاز است. اول اینکه ماشین‌ها با استفاده از رایانه بتوانند اطلاعاتی را از زبان طبیعی انسان به دست آورند، که حوزه فهم زبان طبیعی[2] (NLU) توسط رایانه را شامل می‌شود و دوم اطلاعاتی را به زبان طبیعی منتقل کنند، که در حوزه تولید زبان طبیعی[3] (NLG) در رایانه است [۲].

کاربرد زبان طبیعی شامل دو شاخه نوشتاری و گفتاری می‌شود، یعنی کامپیوتر بتواند هم نوشتار و هم گفتار را پردازش و تحلیل کند. از کاربردهای پردازش زبان طبیعی در شاخه نوشتار می‌توان سرویس‌هایی مانند خطایابی املائی و گرامری در متن، ترجمه ماشینی[4]، استخراج اطلاعات[5] مانند پیدا کردن موضوع یک نوشته، خلاصه‌سازی یک متن، یافتن تشابه یک متن، عقیده کاوی[6] و دستگاه‌های پرسش و پاسخ[7] نام برد و در شاخه گفتاری می‌توان به کاربردهایی مانند تبدیل گفتار به متن[8]، تایپ گفتاری، سامانه‌های اطلاع‌رسانی مانند تلفن گویا، تشخیص گفتار[9] مانند کنترل سیستم از طریق صدا و ترجمه گفتار به گفتار اشاره کرد [۳]. در پردازش نحوی متون زبان فارسی با شاخه نوشتاری سروکار داریم.

انواع زبان‌های طبیعی را از نظر چینش کلمات می‌توان به دو دسته تقسیم کرد. دسته اول زبان‌های با ترتیب مشخص که در آن‌ها واژگان طبق ساختار مشخصی درون جمله قرار می‌گیرند؛ مانند زبان انگلیسی و دسته دوم زبان‌های با آرایش واژگانی آزاد که در جمله امکان

---

1 Natural Language Processing(NLP)
2 Natural Language Understanding(NLU)
3 Natural Language Generating(NLG)
4 Machine Translation
5 Information Extraction
6 Opinion Mining
7 Question Answering Systems
8 Text-to-Speech
9 Speech Recognition

جابه‌جا شدن اجزا و واژگان وجود دارد؛ مانند زبان فارسی [۴].

با افزایش کاربرد رایانه در زندگی بشر و افزایش سندهای متنی الکترونیکی نیاز به پردازش‌های لغوی یا ساخت‌واژه، نحوی و معنایی برای متونی که کاربران با آن‌ها سروکار دارند، احساس می‌شود. یکی از کاربردهای این پردازش‌ها در متون مورد استفاده کاربران، نیاز به خطایاب‌ها و وجود ابزارهایی است که بتوانند این متون را به‌صورت هوشمند ویرایش کنند. خطایابی متون به چهار دسته زیر تقسیم می‌شوند:

- خطایابی لغوی[1]

به تشخیص کلمات دارای غلط املائی در یک متن خطایابی املائی یا لغوی می‌گویند و به ابزاری که غلط‌های املائی یک متن را تشخیص می‌دهد خطایاب املائی[2] گفته می‌شود [۵].

- خطایابی نحوی[3]

برای ایجاد یک متن بدون خطا، توجه صِرف به خطایابی لغوی بدون در نظر گرفتن جایگاه نحوی چندان کافی نیست و نیاز به پرداختن به ابعاد دیگر متن است که در این قسمت به بررسی سطح نحو یا دستور می‌پردازیم. منظور از تحلیل نحو یا دستوری این است که با توجه به قواعد و قوانین مربوط به ترکیب واژه‌ها و ایجاد جملات، ساختار جملات را بررسی کند و مانند یک صافی عمل کند به طوری که اگر جملات با قوانین زبان فارسی همخوانی داشته باشد به آن‌ها اجازه عبور از صافی را بدهد. به‌عنوان نمونه جمله «علی درس را خواندند» از نظر نحوی درست نیست و ساختار درست آن «علی درس را خواند» است [۶].

- خطایابی سبکی[4]

خطایابی سبکی شامل جملاتی است که از نظر نحوی بدون خطا هستند ولی می‌توان آن‌ها را به شکلی بهتر بیان کرد. به عنوان نمونه جمله «به مدرسه علی رفت» از نظر نحوی درست است ولی از نظر سبکی دارای خطا می‌باشد. این خطاها دارای آزادی عمل زیاد هستند و

1 Lexical error detection
2 Spelling checker
3 Grammatical Error Detection
4 Style Error Detection

بیشتر مربوط به ترتیب کلمات در جمله می‌شوند [۶].

- خطایابی معنایی[1]

در این حوزه معنای واژه‌ها و جمله‌های زبان را بررسی و توصیف می‌کنیم. در تحلیل معنایی، هر کلمه یک واحد معنایی به‌حساب می‌آید که یک مجموعه از کلمات به کمک روابط نحوی، معنای یک جمله را تشکیل می‌دهند. به‌عنوان نمونه جمله «کتاب علی را خواند» از نظر نحوی درست است ولی از نظر معنایی نادرست است [۶].

در بین کاربرانِ زبان فارسی، توجه زیادی به رعایت اصول و سبک نوشتار وجود ندارد. به همین دلیل اکثر زبان‌شناسان عقیده دارند، این زبان در صورت عدم توجه کافی دچار نابودی می‌شود. همچنین روزبه‌روز به تعداد اسناد متنی رایانه‌ای اضافه می‌شود و تولید اطلاعات هیچ‌گاه عاری از خطا نیست. پس برای ویرایش لغوی، نحوی و معنایی این متون نیاز به صرف زمان بسیار زیاد و هزینه زیاد است. با توجه به موارد گفته‌شده، اهمیت وجود ابزارهایی به‌طور هوشمند قادرند متون را ویرایش کنند، بیش از پیش احساس می‌شود.

این مقاله می‌کوشد با بهره‌گیری از نظریه دستور زبان وابستگی[2] و با بهره‌گیری از ظرفیتی که فعل برای جمله مشخص می‌کند، ابتدا ساختار جمله را مشخص کند و سپس بر اساس ساختار مشخص‌شده و قوانین نحوی موجود در زبان فارسی، خطاهای نحوی را در متون فارسی تشخیص دهد. پس این مقاله به دنبال ارائه یک مدل برای خطایابی نحوی متون فارسی با استفاده از دستور زبان وابستگی است. مدل ارائه شده، می‌تواند مؤثر برای خطایابی نحوی باشد و از آنجا که در زبان‌هایی با آرایش واژگانی آزاد مثل زبان فارسی پردازش زبان با پیچیدگی‌های زیادی همراه است، بهره‌گیری از دستور زبان وابستگی می‌تواند مفید باشد. با پیاده‌سازی این مدل در ابزارهای هوشمند تحلیل متون می‌توان خطاهای احتمالی نحوی متون را تشخیص داد. در مسیر ساخت یک مدل برای خطایابی نحوی زبان طبیعی با مشکلاتی مواجه هستیم [۶] که تعدادی از این چالش‌ها را در زیر بیان نموده‌ایم.

❖ چندمعنا و چند نقش بودن برخی کلمات **[۶]**

بعضی کلمات در زبان فارسی با وجود شکل نوشتار یکسان دارای چند معنی متفاوت هستند؛

---

1 Semantic Error Detection
2 Dependency Grammar theory

مانند کلمه «شیر». این معانی با توجه به ساختار جمله‌ای که در آن قرار دارند مشخص می‌شوند. همچنین برخی دیگر از کلمات با وجود شکل یکسان، دارای چند معنی و هم چند نقش دستوری در جمله هستند [۶]؛ مانند کلمه «در» که این ویژگی در این کلمات باعث ابهام در متون می‌شود.

❖ حذف به قرینه‌ی کلمات [۶]

این حالت از حذف ممکن است به سه شکل رخ دهد که حالت اول حذف به قرینه‌ی لفظی است که یک یا چند کلمه حذف‌شده در جمله قرار دارند؛ مانند جمله «علی به مدرسه رفت و به دنبالش احمد» که در این جمله کلمه رفت به قرینه لفظی حذف‌شده است [۶].

ممکن است در برخی جملات، کلماتی به قرینه معنوی حذف شوند که برای پی بردن به آن باید به معنا و مفهوم جمله دقت کنیم؛ مانند جمله «آن‌که درسش بهتر، موفق‌تر» که در این جمله کلمات «است» و «می‌شود» بعد از «بهتر» و «موفق‌تر» حذف شده است.

نوع سوم حذف هم حذف به قرینه حضوری است که باید با توجه به موقعیت جمله در شرایط مختلف و متون مختلف کلمه حذف‌شده را حدس زد؛ مانند جمله «... خوب شده است» که با توجه به شرایط ممکن است کلمه حذف‌شده یکی از کلمات «هوا، ظاهرش، اخلاقش و غیره» است.

❖ به کار بردن استعارات و ضرب‌المثل‌ها [۶]

استفاده از استعارات و ضرب‌المثل‌ها باعث بروز مشکل در پردازش متون می‌شود، زیرا معنای این عبارات با ظاهرشان تفاوت بسیاری دارد. همچنین اجزای این عبارات ممکن است به دنبال هم نباشد و یا برخی اجزا در جمله ظاهر نشود که پردازش ساختار این عبارات در جمله را با مشکل روبرو می‌کند. به‌عنوان‌مثال در ضرب‌المثل «هر که بامش بیش، برفش بیش» برخی اجزا ظاهر نشده است [۶].

❖ تشخیص نوع برخی اسامی [۶]

برخی اسامی هستند که ممکن است برعکس ویژگی ظاهری، نوع دسته خود را در تعدد و شمارش تغییر دهند. همچنین در برخی موارد اسامی عام و خاصی وجود دارند که تشخیص آن‌ها در جمله سخت است؛ به عنوان مثال اسم عام «باران» در نقش اسم خاص به‌عنوان اسم دختران به‌کار می‌رود [۶].

❖ نبودن ترتیب در ساختار زبان [۶]

هرچند در زبان فارسی ساختارهایی برای کلمات مانند «فاعل، مفعول، فعل» وجود دارد اما ممکن است در برخی موارد این ساختارها رعایت نشود که همین امر موجب عدم تحلیل سبکی درست در جملات شود. به‌عنوان نمونه جمله «دیروز من کتاب را در مدرسه به مریم دادم» را به چهار شکل مختلف می‌توان نوشت [۷].

❖ حذف کسره‌ی اضافه [۶]

کسره‌ی اضافه در زبان فارسی چند نقش را مشخص می‌کند؛ مانند «صفت و موصوف» و «مضاف و مضاف‌الیه» که دو کلمه را به هم مرتبط می‌کند. همچنین ممکن است بیانگر مالکیت باشد که حذف کسره‌ی اضافه در نوشتار، تشخیص مرزهای عبارات را با مشکل مواجه می‌کند و تشخیص این مرزها توسط ماشین کار سختی است [۶].

❖ عدم تطابق فعل و فاعل در جمله [۶]

ممکن است برخلاف قوانین نظری[1] در یک جمله میان فعل و فاعل از لحاظ تعداد مطابقت وجود نداشته باشد، که این ویژگی پردازش متون را برای ماشین‌ها سخت می‌کند؛ مانند جمله «جناب رئیس تشریف آوردند» که استفاده از فعل جمع برای یک فاعل مفرد به‌منظور احترام است [۶].

❖ معانی و نقش متفاوت در ساختارهای یکسان جملات

در زبان فارسی ممکن است نقش‌های دستوری متفاوت، با حرف‌اضافه مشابه بیایند که پردازش جملات شامل این ویژگی با چالش روبرو می‌شود؛ به‌عنوان نمونه حرف‌اضافه «با» برای نقش‌های توصیف فاعل، بیان ابزار و وسیله، بیان حالت فاعل به کار بیایید؛ مانند «احمد با کلاه آمد»، «احمد با ماشینش آمد» و «احمد با خوشحالی آمد» [۶].

هدف از این مقاله ایجاد یک مدل برای خطایابی نحوی متون فارسی است. در این مقاله مدلی ارائه می‌شود که با کمک ساختار ظرفیتی فعل جمله تمام ویژگی‌های نحوی فعل به دست می‌آید و جمله را از نظر دستوری تحلیل می‌شود. به‌عنوان نمونه پیدا کردن واژه سردرگم در یک جمله که یکی از خطاهای دستوری در جمله است. همچنین به واژه‌هایی که در متن از بعد املائی درست هستند ولی از نظر نحوی در جای درست قرار نگرفته‌اند، پرداخته می‌شود

1 Theoretical Rules

.[۸]

اکثر روش‌هایی که تاکنون ارائه شده، قادر به ایجاد دسته‌هایی مانند اسم، فعل، صفت، ضمیر و غیره برای کلمات هستند که در تحلیل نحوی چندان کارا نیست و حتی برخی واژه‌ها را نیز نمی‌توان در دسته‌ی مشخصی قرار داد.

از نظر طبیب زاده، نخستین پژوهش در مورد دستور زبان وابستگی در زبان فارسی، توسط مهاجر قمی [۹] به‌صورت کتاب انجام شده است. که این کتاب طبق نظریه ظرفیت واژگانی در مورد افعال فارسی نوشته شده است [۱۰].

شهرام احدی [۱۱] در پژوهشی دیگر، کتابی با موضوع متمم‌های نحوی فعل در فارسی برحسب دستور زبان وابستگی نوشته است.

دکتر امید طبیب زاده نیز در پژوهش خود در قالب کتابی، درباره دستور زبان وابستگی با هشت فصل به ظرفیت فعل و ساخت‌های بنیادین جمله در فارسی امروز [۱۰] پرداخته است.

طبق آمار خطاهای املایی در متون حدود ۶۰ درصد از خطاها را شامل می‌شوند و در حدود ۴۰ درصد از خطاها جز خطاهای نحوی و معنایی می‌باشند که این مطلب اهمیت ارائه روشی برای تحلیل نحوی متون فارسی را نشان می‌دهد [۱۲].

۲- دستورزبان مبتنی بر وابستگی

روش مورد استفاده در این مقاله، روش مبتنی بر وابستگی است. از مهم‌ترین روش‌های نمایش دستور زبان، روش‌های نمایش صورت‌گرا[1] است که قابلیت استفاده در مطالعات زبان‌شناسی محاسباتی را دارد [۱۳]. روش مبتنی بر دستور وابستگی[2] از نظریه‌های ساخت‌گرا و صورت‌گرا است که بر اساس ساختار جمله و بدون استفاده از اطلاعات فرامتنی، ساخت‌های زبان را بررسی می‌کند.

در این روش روابط وابستگی بین عناصر هسته[3] (حاکم) و وابسته[4] در زبان را بررسی می‌شود و با توجه به زبان، ساخت‌های نحوی در زبان‌های مختلف را توصیف می‌کند. در نظریه

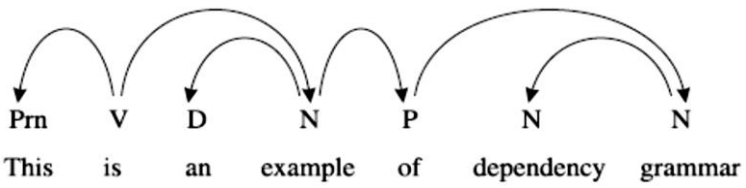
1 Formalist
2 Dependency Grammer
3 Head
4 Dependent

دستور وابستگی، نقش عنصر مرکزی را فعل دارد و دارای جایگاه بالاتری نسبت به عناصر دیگر جمله است که باعث می‌شود ساخت یک جمله با فعل آغاز شود [۱۰].

در روش مبتنی بر وابستگی، یک گروه نحوی به یک کشور با نظام سیاسی فدراتیو شباهت دارد، به بیان دیگر شامل یک حکومت مرزی که همان هسته است و چند ایالت خودگردان یا همان وابسته‌ها است. بزرگ‌ترین گروه نحوی که در دستور زبان وابستگی بررسی می‌شود، جمله است و فعل جمله نقش حکومت مرکزی را دارد و ایالت‌های خودگردان نیز گروه‌های نحوی تشکیل‌دهنده جمله یعنی متمم‌ها و افزوده‌ها هستند. فعل به عنوان هسته جمله بر روابط خارجی میان گروه‌ها نظارت دارد اما روابط داخلی در هر گروه مستقل از فعل است و با توجه به روابط وابستگی بین هسته و وابسته‌ها در خود گروه مشخص می‌شود [۳].

نکته مهم در این روش این است که هر جمله یک فعل مرکزی دارد و بر اساس آن فعل مرکزی و تعداد و نوع متمم‌های اجباری و اختیاری که همان وابسته‌ها هستند، می‌توان ساختار و نحو یک جمله را مشخص کرد. با توجه به مطالب گفته‌شده، ساختار دستور زبان وابستگی برای زبان‌هایی با آرایش واژگان آزاد مناسب است [۳].

برای نمایش ساختار نحوی هر جمله در زبان می‌توان، یک درخت وابستگی که نحوه وابستگی اجزای جمله را مشخص می‌کند، رسم کرد. در شکل زیر درخت وابستگی جمله «This is an example of Dependency grammar» مشخص شده است.

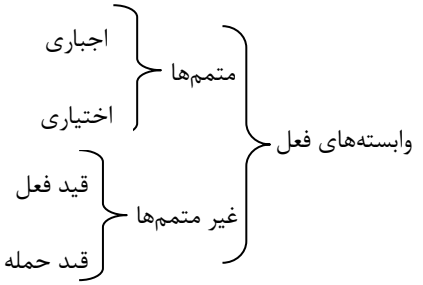


شکل۱: نمونه‌ای از درخت وابستگی [۱۴]

به نظر آقای طبیب‌زاده [۱۰]، وابسته‌های فعل شامل غیرمتمم‌ها(افزوده‌ها)[1] و متمم‌ها[2] هستند

1 Adjunct
2 Complement

و به شکل زیر تقسیم‌بندی می‌شوند.



شکل۲: انواع وابسته فعل [۱۰]

متمم‌ها نیز به دو دسته اجباری[1] و اختیاری[2] تقسیم می‌شود که غالباً به‌صورت اجباری در جمله ظاهر می‌شوند در حالی که غیرمتمم‌ها، همواره به شکل اختیاری هستند. دلیل این امر این است که یک متمم توسط هسته انتخاب می‌شود و رابطه‌ی نزدیکی با هسته دارد در حالی که یک غیر متمم یا افزوده، اطلاعات اضافی را دربردارد و ارتباط نزدیکی با هسته ندارد [۱۵].

۳- مدل پیشنهادی خطایابی نحوی

در این قسمت به دنبال ارائه یک مدل برای خطایابی نحوی و سبکی متون هستیم تا بتوانیم این خطاها را در متون تشخیص دهیم. برای ایجاد یک خطایاب نحوی، ابتدا نیاز است تا واژگان متون از لحاظ ساخت‌واژه بررسی شوند و اشکالات واژگانی آن‌ها برطرف گردد؛ یعنی قبل از ورود به خطایابی نحوی باید تمام لغات از لحاظ املائی، خطایابی شوند و در متون به جای واژگان نادرست، شکل درست آن‌ها نوشته شود. در خطایابی املائی تنها ساخت واژگان را بدون در نظر گرفتن نقش دستوری، مورد تحلیل قرار می‌دهیم.

پس از خطایابی املائی در متون، به کمک مدل پیشنهادی با کمک خطایابی نحوی به بررسی واژه‌هایی می‌پردازیم که از نظر املائی درست هستند، ولی در جمله به صورت اشتباه به کار رفته‌اند.

---

1 Obligatory
2 Optional

همچنین در مدل پیشنهادی به بررسی نوع دیگری از خطاها به نام خطاهای سبکی می‌پردازیم. جملاتی هستند که از نظر نحوی درست نوشته شده‌اند اما ممکن است بتوان همان جملات را از لحاظ ساختار به شکلی بهتر و مناسب‌تر بیان کرد. در این پژوهش علاوه بر بررسی خطاهای نحوی، در قسمتی از مدل پیشنهادی خطاهای سبکی را نیز در متون تشخیص می‌دهیم. در ادامه این مقاله این مدل برای خطایابی نحوی بیان شده است که به توضیح و بررسی آن می‌پردازیم. در این مدل، اسناد متنی به صورت زبان طبیعی وارد می‌شود و پس از انجام مراحل پردازش، خطاهای نحوی در صورت وجود پیدا می‌شوند و نتیجه به کاربر اعلام می‌شود.

شکل۳: مراحل مدل پیشنهادی خطایابی نحوی

همان‌طور که در شکل۳ مشاهده می‌شود، مراحل مدل پیشنهادی برای خطایابی نحوی نمایش داده شده است، که در ادامه به بررسی این مراحل می‌پردازیم:

۳-۱- پیش‌پردازش متن[1]

در اولین مرحله، ابتدا اسناد متون خام را به عنوان ورودی دریافت می‌کند و یک مجموعه از کلمات(یک یا چند کلمه) را به عنوان خروجی برای استفاده در ساختار الگوریتم ارائه می‌دهد.
در پیش‌پردازش، ما به دنبال ساده کردن متون ورودی و کاهش پیچیدگی در آن هستیم، به شکلی که تا حد ممکن مقدار کمی از اطلاعات مفید از بین برود. در متون ورودی باید دقت کرد که علائم نشانه‌گذاری به‌درستی استفاده شده باشند تا پردازش در مراحل بعد به‌درستی انجام پذیرد. در پیش‌پردازش در اولین مرحله، ابتدا واژگان متون را استاندارد و نرمال می‌کنیم و در دومین مرحله، کلمات و جملات را از یکدیگر جدا می‌کنیم.

۳-۲- مشخص کردن برچسب مقوله‌ای، تعداد، شخص و زمان کلمات

با توجه به اینکه در مرحله قبل کلمات از یکدیگر مشخص شده‌اند، در این مرحله به دنبال مشخص کردن نقش نحوی کلمات و خصوصیات نحوی آن‌ها هستیم. ایده‌ای که ما در این قسمت استفاده کرده‌ایم به این صورت است که برای هر کلمه ویژگی‌هایی را به‌دست می‌آوریم. این ویژگی‌ها شامل چهار برچسب مقوله‌ای، برچسب شخص، برچسب تعداد و برچسب زمان هستند که برای تک‌تک کلمات جداگانه به دست می‌آیند.

بدین منظور برای هر کلمه به صورت جداگانه یک قاب[2] تولید می‌کنیم. هر یک از این قاب‌ها دارای شکاف‌هایی[3] جهت پر شدن هر یک از این ویژگی‌ها دارد که برای هر کلمه، متفاوت است. برای پر کردن شکاف‌های قاب کلمات، از یک پایگاه داده پیش‌ساخته که تمام برچسب کلمات در آن ذخیره شده است، استفاده می‌کنیم؛ به عنوان نمونه برچسب‌گذاری کلمه «رفتم» بدین صورت است که ابتدا چهار ویژگی به صورت برچسب مقوله‌ای «فعل»، برچسب شخص «اول شخص»، برچسب تعداد «مفرد» و برچسب زمان «گذشته» در پایگاه داده برای این کلمه استخراج می‌شود، سپس در قاب مشخص خود مطابق جدول زیر، شکاف‌ها پر می‌شوند.

---

1 Text Preprocessing
2 Frame
3 Slots

جدول۱: مثالی از قاب کلمه «رفتم»

نام قاب	رفتم
برچسب مقوله‌ای	فعل
برچسب شخص	اول شخص
برچسب تعداد	مفرد
برچسب زمان	گذشته

ممکن است با توجه به برچسب مقوله‌ای که یک کلمه دارد، ویژگی‌های دیگر یعنی برچسب شخص، تعداد، زمان مقدار مشخص نداشته باشد که شکاف مربوط به آن ویژگی بدون مقدار را با عبارت "NULL" پر می‌کنیم.

۳-۳- تجزیه نحوی وابستگی به کمک درخت وابستگی

در این مرحله، به دنبال تجزیه نحوی جمله به روش وابستگی هستیم. در فصل پیش کلیاتی در مورد دستور زبان وابستگی بیان شد. با برچسب‌گذاری مقوله‌ای کلمات (اسم، صفت و...) نمی‌توان به برچسب‌های سطح بالاتر، نقش کلمات در جمله (فاعل، مفعول و...) دست پیدا کرد. پس با توجه به اینکه در دستور زبان وابستگی، ساخت‌های نحوی جملات در زبان مشخص می‌شود؛ در این مرحله با استفاده از روابط وابستگی که در بین هسته و وابسته وجود دارد، با تعریف یک درخت وابستگی این ساختار را به دست می‌آوریم.

دلیل استفاده از تجزیه نحوی وابستگی، وجود آرایش آزاد کلمات در زبان فارسی است که در تجزیه وابستگی تحمل‌پذیری بالاتری دارد. با توجه به اینکه در تجزیه وابستگی، مهم‌ترین جزء یک جمله، فعل آن است؛ در این پژوهش با کمک ظرفیت فعل در نظریه دستور وابستگی، به دنبال خوش‌ساخت کردن جمله هستیم.

اولین گام در به دست آوردن ساختار پیوستگی، تجزیه نحوی وابستگی است. یکی از ابزار موجود در زبان فارسی برای تجزیه نحوی وابستگی، ابزار «هضم» است که در این پژوهش از این ابزار برای تجزیه نحوی استفاده می‌شود. ابزار هضم عمل تجزیه نحوی را به کمک این پیکره دادگان [۱۶] انجام می‌دهد.

۳-۴- مشخص کردن نقش نحوی کلمات از طریق قوانین نحوی

در این مرحله، ابتدا قوانینی برای تشخیص نقش وابسته‌های نحوی فعل در جمله تعریف می‌کنیم و سپس بدون توجه به ترتیب وابسته‌های فعل در جمله، به تمام کلمات بر اساس این قوانین از پیش تعریف‌شده، برچسب نحوی می‌دهیم. در این پژوهش از قوانین زیر استفاده می‌کنیم [۱۷].

۱. یک اسم یا گروه اسمی که قبل از حرف‌اضافه «را» آمده باشد، از نظر نحوی دارای نقش «مفعول» است؛ مانند «علی **کتاب** را خواند.»

۲. به اسمی که بلافاصله بعد از «حرف‌اضافه»؛ مانند «به، از، در، با، برای، مثل و مانند این‌ها» آمده باشد، از نظر نحوی یک وابسته «مفعول حرف‌اضافه‌ای» است؛ مانند «علی با **اتومبیل** رفت.»

۳. اگر فعل جمله مرکب باشد و یک متمم به کمک یک «حرف‌اضافه» به جزء غیر فعلی فعل مرکب، پیوند خورده باشد، در این صورت به این وابسته‌ی نحوی فعل مرکب «مفعول نشانه اضافه‌ای» می‌گویند؛ مانند «دانشگاه به من اجازه‌ی تدریس داد.» حروف دیگری هم به نام «حرف ربط»؛ مانند «و، یا، ولی، زیرا» وجود دارند، که نباید آن‌ها را با «حرف‌اضافه» اشتباه گرفت.

۴. همان‌طور که در جدول ۳-۸ آمده، محل وابسته نحوی «فاعل» در ساختار جمله، ابتدای جمله است و اگر «فاعل» در جایگاه‌های دیگری به جز ابتدای جمله آورده شود، خطای سبکی می‌باشد. همان‌طور که در ابتدای فصل اشاره شد، خطایابی سبکی مربوط به این قسمت از مدل است.

۵. محل قرارگیری «هسته» در ساختار جمله، در انتهای جمله است. به عبارتی چون در دستور زبان وابستگی، هسته جمله «فعل» است، پس باید همواره انتهای جمله؛ «فعل» باشد.

۶. در جملات، «مسندها» برچسب مقوله‌ای «اسم یا صفت» دارند و کلمات با برچسب «قید» در جملات، نمی‌توانند نقش نحوی مسند را در ساختار جمله داشته باشند؛ مانند «محمد خوشحال است.»

۷. ساختار یک جمله باید بر اساس ساختار ظرفیتی فعل جمله باشد، در غیر این صورت دارای خطای نحوی است.

۳-۵- خطایابی بر اساس معیارهای خطایابی

در این مرحله جملات را بر اساس پنج معیار خطایابی، بررسی می‌کنیم:

- معیار اول خطایابی: بررسی وجود فعل در جمله

در مرحله بعد از برچسب‌گذاری مقوله‌ای، جمله با توجه به اولین معیار خطایابی نحوی یعنی «وجود فعل در جمله»، بررسی و تحلیل می‌شود. در این مرحله برای بررسی وجود فعل در جمله اگر در مجموعه برچسب‌های تخصیص داده‌شده به کلمات، برچسب فعل مشاهده نشد، یک خطای نحوی «کمبود فعل در جمله» گرفته می‌شود و در غیر این صورت، اگر در مجموعه برچسب‌های تخصیص داده‌شده، فعل مشاهده شد، جمله به مرحله بعد و بررسی با معیارهای بیشتر می‌رود.

به عنوان نمونه، در بررسی عبارت «او در مسابقات جهانی رتبه اول را» ابتدا به تک‌تک کلمات آن، برچسب مقوله‌ای نسبت می‌دهیم. تحلیل صرفی این عبارت برای به دست آوردن برچسب‌ها به‌صورت « او/PRO، در/P، مسابقات/Ne، جهانی/AJ، رتبه/Ne، اول/NUM، را/POSTP» استخراج می‌شود. با توجه به تحلیل صرفی این عبارت مشاهده می‌شود که در آن برچسب فعل وجود ندارد. به همین دلیل یک خطای نحوی «کمبود فعل در جمله» گرفته و به کاربر اعلام می‌شود.

- معیار دوم خطایابی: تطابق بین برچسب زمان فعل قید زمان[1](در صورت وجود) در جمله

قید زمان، یک کلمه یا گروهی از کلمات است که زمان انجام فعل را بیان می‌کند؛ مانند امروز، شب، سحرگاه، زود، دیر، امسال، دیروز، اکنون، دیشب و مانند این‌ها [۱۸]. با توجه به اینکه در مرحله برچسب‌گذاری، برچسب زمان فعل و برچسب زمان قید را مشخص کرده‌ایم، در این مرحله برچسب زمان قید زمان را با برچسب زمان فعل جمله مقایسه می‌کنیم. در صورتی که برچسب زمان فعل در جمله با برچسب زمان قید زمان در حالت‌های «گذشته» و «آینده» مطابقت نداشته باشد، یک خطای نحوی تحت عنوان «عدم تطابق قید زمان با زمان فعل» تشخیص می‌دهد و این خطا به کاربر اعلام می‌شود. قید زمان «حال»، ممکن است همراه افعال با هر زمانی به کار برده شود. همچنین قید زمان «آینده» نیز ممکن است با فعل با زمان

1 Adverb Of Time

«حال» و «آینده» به کار برود که این مورد را نیز در بررسی‌ها، در نظر می‌گیریم.

به‌عنوان نمونه در جمله‌ی «موسسه آموزش عالی سپاهان پارسال به سطح یک کشوری می‌رسد.» برچسب زمان فعل «می‌رسد»، **حال** است در حالی که در جمله دارای قید زمان «پارسال» دارای برچسب زمان گذشته است و این یک خطای نحوی در جمله است.

- معیار سوم خطایابی نحوی: بررسی ساختار وابستگی[1] جمله به کمک قوانین نحوی

در این مرحله، ساختار نحوی به‌دست‌آمده از جمله در مراحل قبل را به کمک قوانین نحوی بررسی می‌کنیم. در صورتی که یکی از قوانین در نقش نحوی کلمات رعایت نشده باشد، خطای مربوطه اعلام شود. در غیر این صورت، اگر تمام قوانین در ساختار نحوی جمله صادق بودند، برای بررسی بیشتر به مراحل بعد می‌رویم؛ به عنوان نمونه جمله «او امروز است.» و «به مدرسه علی رفت.» را با این روش، بررسی می‌کنیم. جمله اول، در خطایابی نحوی پس از عبور از معیار اول و دوم خطایابی، در این مرحله طبق شکل زیر ابتدا ساختار وابستگی آن به دست می‌آید و وابسته‌های فعل، نقش نحوی می‌گیرند.

شکل۴: درخت وابستگی یک جمله نمونه با خطای نحوی

- معیار چهارم خطایابی: شکل صحیح مسند در جمله

این مرحله از خطایابی نحوی، فقط در جملات اسنادی استفاده می‌شود. «جملات اسنادی» به جملاتی گفته می‌شود که در آن حالتی به «مسند الیه» (فاعل) نسبت داده می‌شود. به این حالت انتساب داده‌شده که اسم یا صفت است، «مسند» گفته می‌شود. در

1 Dependency Structure

جملات اسنادی از افعال ربطی[1](اسنادی) «است، بود، شد، گشت، گردید و مانند این‌ها» استفاده می‌شود؛ به‌عنوان‌مثال جمله «علی دلیر است.»، یک جمله‌ی اسنادی است که حالت «دلیر» به عنوان مسند به «علی»، نسبت داده شده است.

در جملاتی اسنادی، اگر مسند برچسب تعداد «جمع» داشته باشد، یک خطای نحوی رخ داده و تحت عنوان «عدم تطابق مسند جمع با مسندالیه مفرد» شناخته و در لیست خطاها ذخیره می‌شود؛ به عنوان مثال در جمله «ما دلیران هستیم.» کلمه «دلیران» به عنوان مسند جمله، با مسندالیه یعنی کلمه «ما» تطابق ندارد.

- معیار پنجم خطایابی: تطابق بین فاعل و فعل از نظر برچسب شخص و برچسب تعداد

آخرین مرحله در خطایابی در مدل پیشنهادی، مقایسه بین «فاعل» و «فعل جمله» از نظر برچسب «شخص و تعداد» است و در صورتی که در یکی از این دو برچسب، بین فاعل و فعل جمله یکسان نباشد، در جمله از نظر نحوی خطا تشخیص داده می‌شود و در لیست خطاهای نحوی درج می‌شود. به عنوان مثال در جمله «تو به مدرسه رفتند.»، فعل «رفتند» یک فعل با برچسب شخص «سوم شخص» و برچسب تعداد «جمع» است، در حالی که «تو»، یک ضمیر با برچسب شخص «دوم‌شخص» و برچسب تعداد «مفرد» است. در این مثال، هم برچسب شخص و هم برچسب تعداد با یکدیگر تطابق ندارد و یک خطای نحوی شناخته می‌شود.

۳-۵- مرحله اعلام نتیجه

در این بخش، با توجه به بررسی جمله در معیارهای دوم، سوم، چهارم و پنجم در خطایابی نحوی، اگر جمله از نظر هیچ یک از معیارها، دارای خطا نبود، «صحیح بودن جمله» را اعلام می‌کند؛ در غیر این صورت در صورت وجود خطا در هر یک از این معیارها، آن خطا به عنوان یک خطای نحوی اعلام می‌گردد.

1 Related Verbs

۴- نتایج و ارزیابی

برای ارزیابی عملکرد مدل گفته‌شده، بهترین راه مقایسه با کارهای پیشین درزمینه‌ی خطایابی نحوی است، به صورتی که بر اساس پارامترهایی مشخص، در مقایسه با روش‌های دیگر، مزایا و معایب آن سنجیده و بررسی شوند. در زبان‌های غیرفارسی، کارهایی در این زمینه انجام شده است؛ به عنوان مثال ناتسون[1] با کمک همکارانش مقاله‌ای برای خطایابی نحوی، در زبان سوئدی ارائه داد که حدود ۳۰ درصد از خطاهای نحوی را تشخیص می‌داد [۶].

با توجه به ساختار و ویژگی‌های زبان فارسی، مانند ساختار واژگان آزاد[2] و تفاوت‌هایی اساسی که این زبان با زبان‌های دیگر دارد و اینکه خطایابی نحوی در آن زبان‌ها، جنبه‌ای از جملات را شامل شود که در زبان فارسی حائز اهمیت نیست، نمی‌توان مقایسه درستی بین روش این پژوهش و روش‌های زبان‌های دیگر ارائه کرد. همچنین در زبان فارسی، برخلاف زبان‌های دیگر کار چندانی در رابطه با تجزیه متون و خطایابی نحوی صورت نگرفته است و این نیز دلیلی برای عدم استفاده از مقایسه برای ارزیابی است.

بدین منظور برای ارزیابی مدل پیشنهاد شده از معیارهای مشخص شده‌ای استفاده می‌کنیم. این نمونه‌ها شامل دو مجموعه از جملات هستند.

- مجموعه جملات اول

این مجموعه جملات انتخاب‌شده، شامل ۲۰۵ خبر از سایت خبری تابناک هستند که جملاتی بدون خطای نحوی هستند و در ارزیابی آن‌ها به دنبال این هستیم که «آیا ممکن است مدل پیشنهادی به اشتباه در برخی از این جملات درست خطای نحوی تشخیص دهد.»

- مجموعه جملات دوم

این مجموعه جملات شامل ۲۰۰ جمله از ۴۰ نفر است که با تشریح هر یک شاخص‌های خطایابی نحوی از آنان خواسته شد تا هر یک پنج جمله در رابطه با این شاخص‌ها بیان کنند و همچنین بر این سعی داشته باشند، جملاتی بگویند که دارای خطای نحوی در این پنج شاخص

---

1 Nutson
2 Free Word Structure

باشند و مدل پیشنهادی ما از خطایابی آن‌ها ناتوان باشد، به عبارتی ما در ارزیابی به دنبال جواب؛ «آیا ممکن است در بررسی مدل پیشنهادی به اشتباه، جملاتی که دارای خطای نحوی هستند، بدون خطا در نظر گرفته شوند.»

سپس این جملات را بر اساس پنج شاخص زیر که در ساختار مدل مورد بررسی قرار می‌گرفت، مورد ارزیابی قرار می‌دهیم.

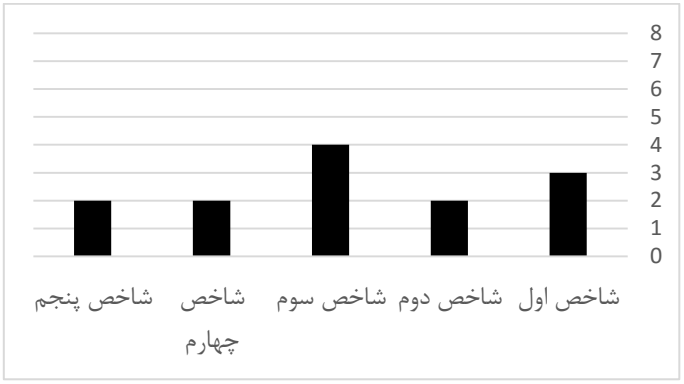
جدول۲: شاخص‌های ارزیابی در خطایابی نحوی

شاخص	بررسی نحوی انجام‌شده
شاخص اول ارزیابی	بررسی وجود فعل در جمله
شاخص دوم ارزیابی	بررسی تطابق بین برچسب زمان فعل با قید زمان
شاخص سوم ارزیابی	بررسی وجود خطا در ساختار جملات
شاخص چهارم ارزیابی	بررسی تطابق برچسب تعداد مسند و مسندالیه
شاخص پنجم ارزیابی	بررسی تطابق بین برچسب شخص و تعداد فاعل با فعل

۴-۱- نتایج حاصل از ارزیابی مجموعه جملات اول

با اعمال روش پیشنهادی خطایابی نحوی بر روی ۲۰۰ جمله از اخبار سایت تابناک، به نتایج قابل‌توجهی دست پیدا کردیم. پس از این آزمایش، روش پیشنهادی ۱۳ جمله را حاوی خطای نحوی تشخیص می‌دهد. به بیان دیگر، این روش به طور اشتباه ۱۳ خبر از این اخباری که درست هستند را دارای خطای نحوی تشخیص داده است. با توجه به اینکه در روش این پژوهش تک‌تک جملات در پنج شاخص مورد بررسی قرار می‌گیرد و در صورت مشاهده خطا، خطای مربوطه به کاربر اعلام می‌شود، پس از بررسی این ۶ درصد از اخبار که به اشتباه خطایابی شده‌اند، پراکندگی[1] آن‌ها در هر یک از شاخص‌های ارزیابی به صورت زیر است.

1 Dispersion



نمودار۱: میزان پراکندگی خطایابی اشتباه مجموعه جملات اول

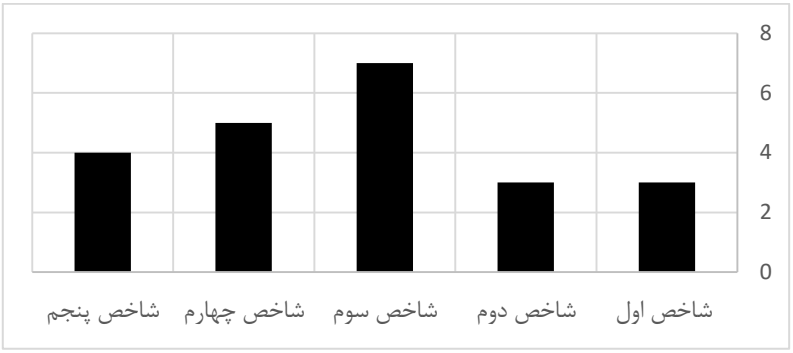
۴-۲- نتایج حاصل از ارزیابی مجموعه جملات دوم

در این قسمت، به بررسی نتایج حاصل از خطایابی نحوی مدل پیشنهادی بر روی ۲۰۰ جمله آزمایشی می‌پردازیم. بیشتر افرادی که جملات را گفتند، سعی داشتند تا از جملات غلط استفاده کنند تا کارایی مدل ما را مورد ارزیابی قرار دهند. پس از بررسی جملات توسط روش این پژوهش، جملات آن‌ها در دو رده زیر تقسیم‌بندی شدند.

جدول۳: تقسیم‌بندی جملات مجموعه جملات دوم

جملات دارای خطای نحوی	جملات بدون خطای نحوی
۱۲۸	۷۲

حال به ارزیابی عملکرد روش خطایابی در این جملات می‌پردازیم. پس از بررسی خطایابی نحوی مجموعه جملات دوم در مدل پیشنهادی، این مدل در ۱۱ درصد از جملات، تصمیم اشتباه گرفته است و به اشتباه جملاتی را که دارای خطا بودند را بدون خطا در نظر گرفته است. در نمودار زیر، میزان پراکندگی خطایابی اشتباه، در هر یک از شاخص‌ها مشخص شده است.



نمودار۲: میزان پراکندگی خطا در شاخص‌ها

۴-۳- ارزیابی مدل پیشنهادی با معیارهای ارزیابی

برای ارزیابی عملکرد روش‌های هوشمند، از معیارهای ارزیابی جهت بررسی نمونه‌های آزمایشی استفاده می‌شود. برای به دست آوردن معیارهای ارزیابی، نیاز است تا ابتدا برای این نمونه‌ها ماتریسی به نام «ماتریس درهم‌ریختگی»[1] داشته باشیم. برای داشتن ماتریس درهم‌ریختگی باید نتایج نمونه‌های ما توسط دو شاخص، دسته‌بندی شوند. اگر بتوانیم نمونه‌ای آزمایشی را به دو گروه مثبت و منفی تقسیم‌بندی کنیم، آنگاه می‌توان آن‌ها را از طریق این معیارها مورد ارزیابی قرار داد [۱۹]. در روش این پژوهش نیز با توجه به مجموعه جملات ورودی در مرحله قبل، باید ماتریس درهم‌ریختگی مطابق جدول زیر ایجاد کنیم تا به کمک آن بتوانیم روش را به کمک معیارهای ارزیابی مورد بررسی قرار دهیم.

جدول۴: ماتریس درهم‌ریختگی با توجه به مجموعه جملات آزمایشی

		نتیجه مدل ارائه‌شده	
		خطایابی نحوی(مثبت)	عدم‌خطایابی نحوی (منفی)
واقعیت	جمله دارای خطای نحوی(مثبت)	۱۰۶	۲۲
	جمله فاقد خطای نحوی (منفی)	۱۳	۲۶۴

برای ارزیابی مدل ارائه‌شده در این پژوهش به کمک معیارهای گفته‌شده، ابتدا به

1 Confusion Matrix

اطلاعاتی از نتایج به دست آمده نیاز داریم که این اطلاعات از ماتریس درهم‌ریختگی ارزیابی به دست می‌آیند. با توجه به دو مجموعه جملات، اطلاعات کلی به‌صورت زیر خواهند.

جدول۵: اطلاعات حاصل از دو مجموعه جملات آزمایشی

کل جملات دو دسته از جملات	۴۰۵
تعداد خطایابی نحوی درست[1]	۱۰۶
تعداد خطایابی نحوی اشتباه[2]	۱۳
تعداد عدم خطایابی نحوی درست[3]	۲۶۴
تعداد عدم خطایابی اشتباه[4]	۲۲

حال با کامل شدن ماتریس درهم‌ریختگی، به دنبال به دست آوردن معیارهای ارزیابی هستیم. پس از به دست آوردن مقادیر معیارهای ارزیابی با تبدیل آن‌ها به درصد، جدولی به صورت زیر برای بیان درصد هر یک از این معیارها خواهیم داشت.

جدول۶: نتایج حاصل از معیارهای ارزیابی طبق جملات آزمایشی

معیارهای ارزیابی	نتایج بدست‌آمده در مدل پیشنهادی
ACC	٪۹۱
ERR	٪۹
PPV	٪۸۹
TPR	٪۸۳
NPV	٪۹۲
NPV	٪۹۵
$F-Measure$	٪۸۶

با توجه نتایج به دست آمده از درصد معیارهای ارزیابی در جدول۶ میزان دقت کل در

1 True Positives
2 False Positives
3 True Negatives
4 False Negatives

مجموعه جملات آزمایشی، برابر با ۹۱٪ است. پس می‌توان گفت که مدل پیشنهادشده در این پژوهش با رویکرد مبتنی بر دستور زبان وابستگی برای خطایابی نحوی، دارای دقت بالایی است. افزایش دقت، به این دلیل است که خطایابی نحوی در چند معیار انجام شده و سعی شده تمامی ابعاد نحوی جملات در زبان فارسی در این مدل، لحاظ شود. همچنین میزان اشتباه در تشخیص خطاها، برابر با ۹٪ است که در قسمت‌های قبل به دلیل این اشتباهات در خطایابی پرداختیم. نکته قابل‌توجه در تحلیل نتایج این است که این مدل در ۸۹٪ از خطایابی‌های نحوی که انجام داده، به درستی عمل انجام داده است. همچنین موفق به پیدا کردن ۸۳٪ از خطاهای نحوی در جملات شده است. نکته دیگر این است که ۸٪ از خطاهای نحوی، به اشتباه انجام نشده است. از طرفی ۵٪ از جملات بدون خطای نحوی، به اشتباه خطادار شناخته شده است.

۵- جمع‌بندی و نتیجه‌گیری

در کشورهای ایران، افغانستان، تاجیکستان و کشورهای اطراف بیش از ۱۰۰ میلیون نفر، زبان فارسی را به عنوان زبان اول و مادری می‌شناسند و با آن صحبت می‌کنند. با این حال با افزایش زندگی ماشینی و ورود کامپیوتر در زندگی بشر و همچنین افزایش سندهای متنی، ایجاد یک خطایاب نحوی جهت بررسی نحوی متون در زبان فارسی، اهمیت ویژه‌ای پیدا کرده است. دلیل این اهمیت این است که در زبان فارسی بسیاری از کاربران، دقت کافی در رعایت اصول نوشتاری نمی‌کنند و این زبان در خطر نابودی است.

استفاده از خطایابی نحوی می‌تواند کمک زیادی در رفع این مشکل کند. به طور کلی، برای افزایش فهم زبان نوشتار و گفتار در ماشین و همچنین ترجمه ماشینی، به یک ابزار قوی برای خطایابی نحوی نیاز داریم. در خطایابی نحوی اساس و پایه کار، درک نحوی زبان طبیعی توسط کامپیوتر است. در این قسمت به کاربردهای دیگر این مدل اشاره می‌کنیم.

- سیستم‌های ترجمه ماشینی

همان‌طور که در قبل گفتیم، ترجمه ماشینی به استفاده از کامپیوتر برای ترجمه یک متن از یک زبان طبیعی به زبان دیگر گفته می‌شود. از آن جا که برای ترجمه متون به زبان مقصد، باید ساختار نحوی و قوانین نحوی آن زبان را بدانیم، وجود یک سیستم که بتواند به تولید

جملات با معنی و بدون خطای نحوی در ترجمه ماشینی کمک کند، حائز اهمیت است.

-خطایاب برای زبان‌های دیگر

مدل پیشنهادی در این پژوهش برای خطایابی نحوی برای زبان‌هایی با آرایش واژگان آزاد در نظر گرفته شده است که به این دلیل می‌توان از ایده این پژوهش و سیستمی که پیشنهاد شد را برای زبان‌های طبیعی دیگر و تعمیم داد و پیاده‌سازی کرد. برای این کار باید دستور زبان وابستگی آن زبان طبیعی، در نظر گرفته شود و پایگاه داده برچسب‌های کلمات در آن زبان، ایجاد شود.

-سیستم دسته‌بندی جملات بر اساس قوانین دستوری

همچنین می‌توان جملات و کلمات را بر اساس قوانین دستوری و برچسب‌های نحوی به دسته‌های مشخص، تقسیم‌بندی کرد؛ به عنوان نمونه در یک متن کلمات با «برچسب تعداد جمع» را در یک دسته قرار داد یا افعال با «برچسب زمان گذشته» را در دسته مخصوص به خود قرار داد. این کار برای کارهای آماری در متون و تقسیم‌بندی‌های نحوی متون، می‌تواند کاربرد داشته باشد.

-خطایابی تخصصی

در خطایابی تخصصی می‌توان این امکان را ایجاد کرد که علاوه بر خطایابی نحوی از قوانین معنایی نیز استفاده شود تا بتوان اشتباهات خطایابی نحوی را کم و دقت آن را افزایش داد.

-آموزش زبان فارسی در سطح بین‌الملل

همچنین مدل گفته شده، می‌تواند کمک شایانی در جهت آموزش زبان فارسی به غیر فارسی زبانان در زمینه‌ی دستور و گرامر زبان فارسی داشته باشد و روند یادگیری این زبان را تسهیل کند.

در این پژوهش، به دنبال ارائه مدلی برای بررسی نحوی متون و پیدا کردن خطاهای نحوی و سبکی بودیم. فازهای دیگر از خطایابی، یعنی خطایابی املائی و خطایابی معنایی در

کنار خطایابی نحوی و سبکی به ارائه متون فارسی عاری از خطا منجر می‌شود که امیدواریم در آینده این امر محقق شود.

در نهایت امید است که این پژوهش، راه گشا و مؤثر برای پردازش زبان فارسی به خصوص در زمینه پردازش نحوی متون فارسی و خطایابی نحوی و همچنین یاری کننده علاقه‌مندان پژوهش در این حوزه باشد.

منابع

[۱] ویکی پدیا. ۲۰۱۷. "پردازش زبان طبیعی". قابل دسترس از:
http://fa.wikipedia.org/wiki/پردازش_زبان_طبیعی

[2] Kelleher and Genabith. 2005. "Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. in Artificial Intelligence. pp. 62-102

[3] D. Jurafsky, J. H. Martin. 2009. "Speech and Language Processing: An Introduction to Natural Language ProcessingComputational Linguistics, and Speech Recognition. 2 ed. New Jersey: Pearson: Prentice Hall.

[4] Comrie. 1989. "Language Universals and Linguistic Typology". Chicago: Univesity of Chicago Press

[۵] شورای‌عالی‌اطلاع‌رسانی. ۱۳۸۸. "مقدمه‌ای بر طراحی و ایجاد خطایاب املایی صرفی زبان فارسی".طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی دانشگاه علم و صنعت تهران.

[۶] شورای‌عالی‌اطلاع‌رسانی. ۱۳۸۸. "استخراج نیازمندی‌های ابزار خطایاب املایی در لایه نحو زبان فارسی به پیکره‌های فارسی مورد نیاز". طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی دانشگاه علم و صنعت تهران.

[۷] شمس فرد م. ۱۳۸۴. پردازش متون فارسی؛دستاوردهای گذشته؛ چالش‌های پیش رو. دومین کارگاه پژوهشی زبان فارسی و رایانه، تهران.

[8] Jin Hu Huang and David Powers. 2001. "Large Scale Experiments on Correction of Confused Words". Proceeding of the 24th Australasian conference on Computer Science.

[9] Mohadjer-Ghomi S. 1978. "Eine Kontrastive Untersuchung" der Satzbauplane in Deutchen und Persishen. Kirchzarten: Burg-Verlag

[۱۰] طبیب‌زاده امید. ۱۳۸۵. "ظرفیت فعل و ساخت‌های بنیادین جمله در زبان فارسی امروز". تهران: نشر مرکز.

[11] Ahadi S. 2001. "Verberganzungen und zusammengesetzte Verben im Persischen". Wisebaden: Verla

[12] Hussain, Dr.Sarmad, Naseem, Tahira. 2006. "Spell checking. CRULP, NUCES, PAKISTAN

[13] Quirk R. 1985. "A Comprehensive Grammar of the English Language". London and New York: Longman

[14] Liu C, Wang H, Mcclean S, Liu J, Wu S . 2007. "Syntactic information retrieva". Proceeding of the IEEE International Conference on Granular Computing

[۱۵] طبیب‌زاده امید. ۱۳۸۳. "وابسته‌های فعل در زبان فارسی بر اساس نظریه‌ی وابستگی". نامه‌ی فرهنگستان (ویژه نامه‌ی دستور)، شماره ۵. صفحات ۱۳-۲۸.

[16] Rasoli M.S, Moloodi A, Kouhestani M, Minaei-bidgoli B. 2011. "A Syntactic Valency Lexicon for Persian Verbs: The First Steps to words Persian Dependency Treebank". 5th Language & Technology Conference(LTC) :Human Language Technologies as a Challenge for Computer Science and Linguistics

[۱۷] دبیرمقدم محمد. ۱۳۸۳. "زبان فارسی و نظریه‌های زبانی: در جستجوی چارچوبی برای تدوین دستور جامع زبان فارسی". دستور.

[۱۸] انوری حسن. ۱۳۹۰. "دستور زبان فارسی". چاپ اول اصفهان.

[19] Kohavi R, Provost F. 1998. "On Applied Research in Machine Learning In Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process". Columbia University, New York. Vol 27

پیکرهٔ زبان‌آموز «سلام فارسی»
معرفی الگوی دسته‌بندی و تعیین خطاها، مجموعهٔ برچسب و ابزار برچسب‌دهی

سعید صفری*

چکیده

«سلام فارسی»، پیکرهٔ زبان‌آموز فارسی است که به منظور جمع‌آوری تولیدات نوشتاری و گفتاری فارسی‌آموزان خارجی و نشانه‌گذاری خطاهای زبانی ایشان طراحی و تشکیل شده است. این پیکره قادر به ارایه گزارش‌های متعددی از جمله تعیین فراوانی و نوع خطاهای فارسی‌آموزان و دسته‌بندی آن بر اساس معیارها و فراداده‌های گوناگون از جمله کشور، نوع زبان اول و جنسیت است. هدف از تشکیل این پیکره، بررسی و تحلیل زبانشناختی خطاهای فارسی‌آموزی است که می‌تواند نقشهٔ راهی برای تعیین بهتر راهبردهای آموزشی و تهیه و تولید محتوای درسی هدفمند و متناسب با نیاز فارسی‌آموزان ترسیم کند. در این مقاله الگوی تعیین خطاها، تهیهٔ مجموعه برچسب و شیوه برچسب‌زنی داده‌های پیکره با استفاده از ابزار برچسب‌دهی و ویرایش خطاها معرفی می‌شوند.

واژه‌های کلیدی: پیکرهٔ زبان‌آموز، زبان‌شناسی پیکره‌ای، تعیین خطاهای فارسی‌آموزی، آموزش زبان فارسی به غیر فارسی زبانان.

۱. مقدمه

امروزه استفاده از پیکره‌ها در بسیاری از پژوهش‌های زبانی امری ضروری و اجتناب‌ناپذیر است؛ پیکره‌ها امکان دسترسی به داده‌ها و نمونه‌های عینی زبانی را فراهم می‌کنند و تحلیل و توصیف مسائل زبانی با اتکا به گزارش‌ها و تحلیل‌های پیکره دقیق‌تر صورت می‌گیرد. در حوزهٔ پژوهش‌های زبان دوم/خارجی‌آموزی، پیکره‌های زبان‌آموز که تولیدات نوشتاری یا گفتاری زبان‌آموزان را جمع‌آوری و نشانه‌گذاری می‌کنند، امکان مطالعه ویژگی‌های زبان میانی را فراهم

* دانشگاه بلگراد، saeed.safari@fil.bg.ac.rs

می‌آورند. گرانجر (۲۵۹: ۲۰۰۸)[۱] با اشاره به اهمیت پیکره‌های زبان‌آموز برای ارایه توصیف بهتر از زبان میانی و عوامل تاثیرگذار بر آن، به‌کارگیری این پیکره‌ها را «برای دستیابی به نظریهٔ یادگیری زبان دوم» حائز اهمیت می‌داند. هرچند پیشینهٔ تشکیل و تولید پیکره‌های زبان‌آموز به سالهای پایانی قرن بیستم می‌رسد (کالیس و پاکوت، ۲۰۱۵) [۲]، اما با افزایش اقبال پژوهشگران و متخصصان آموزش زبان دوم/خارجی به استفاده از گزارش‌ها و تحلیل‌های مبتنی بر پیکره‌های زبان‌آموز، حوزه مطالعاتی و بینارشته‌ای جدیدی با عنوان «پژوهش‌های پیکرهٔ زبان‌آموز»[1] شکل گرفته است که به نقل از گرانجر و همکاران (۲۰۱۵:۳) [۳] «حلقهٔ ارتباطی بین زبان‌شناسی پیکره‌ای، فراگیری زبان دوم، روش تدریس زبان خارجی و پردازش زبان طبیعی را برقرار می‌کند.». با توسعه و افزایش کمّی پیکره‌های زبان‌آموز، دسته‌بندی‌های متعددی بر اساس معیارهای مختلف صورت گرفته که یکی از مهمترین آنها، دسته‌بندی بر اساس «نوع نشانه‌گذاری داده‌های پیکره» است. در این دسته‌بندی، پیکره‌هایی که برچسب خطاهای زبانی نشانه‌گذاری می‌شوند، و از این پس در این مقاله آنها را «پیکره‌های خطایابی زبان‌آموز» می‌نامیم، برای پژوهشگران حوزهٔ فراگیری زبان دوم، متخصصان آموزش زبان، طراحان برنامه‌درسی و مولفان درسنامه‌های آموزشی اهمیت بیشتری دارند زیرا با تعیین نوع و فراوانی خطاهای زبانی مربوط به گروه‌های مختلف زبان‌آموزان، علاوه بر امکان مطالعه نظری شیوه‌های فراگیری زبان دوم، امکان برنامه‌ریزی و تهیه راهبردهای آموزشی مناسب و همچنین تولید محتوا و منابع درسی هدفمند فراهم می‌شود (نسلهاف، ۲۰۰۴) [۴]. در این مقاله، نخست توضیح مختصری دربار، پروژهٔ تشکیل و تولید پیکرهٔ زبان‌آموز «سلام فارسی» و اهداف آن ارایه می‌شود و در بخشهای بعدی، الگوی دسته‌بندی و تعیین خطاها، تشکیل مجموعه برچسب‌ها و شیوه برچسب‌زنی پیکره معرفی می‌شوند. در بخش پایانی و به صورت خلاصه، نتایج کلی از گزارش‌های این پیکره درباره فراوانی و نوع خطاهای فارسی‌آموزان صرب ارایه می‌گردد.

۲. پیکرهٔ زبان‌آموز «سلام فارسی»

«سلام فارسی»[2]، پیکرهٔ زبان‌آموزی است که با هدف جمع‌آوری و نشانه‌گذاری خطاهای

1 Learner Corpus Research (LCR)
2 The Salam Farsi Learner Corpus (SFLC)

زبانی فارسی‌آموزان در دانشکدهٔ ادبیات و زبان‌های خارجی دانشگاه بلگراد[1] طراحی و تولید شده است. این پیکره در فاز نخست به جمع‌آوری و برچسب‌زنی خطاها در تولیدات نوشتاری فارسی‌آموزان صرب پرداخته است و طرح توسعه در دو فاز پژوهشی دیگر را در دست اجرا دارد: در فاز دوم، تولیدات نوشتاری فارسی‌آموزان با پیشینهٔ زبان‌های اسلاوی (در منطقهٔ بالکان و اروپای شرقی) و در فاز سوم، تولیدات فارسی‌آموزان با پیشینهٔ زبانی گوناگون را جمع‌آوری و نشانه‌گذاری خواهد نمود. گردآوری و استفاده از تولیدات نوشتاری فارسی‌آموزان در پیکره با کسب موافقت کتبی از منابع تامین‌کنندهٔ داده‌ها و طبق توافق پژوهشی بین دانشگاه بلگراد و دانشگاه‌ها و موسسات آموزشی مرتبط صورت می‌گیرد.

در این مقاله اطلاعات مربوط به فاز نخست و داده‌های جمع‌آوری شده از فارسی‌آموزان صرب ارایه می‌شود. در تشکیل این پیکره، معیارهایی برای طراحی آن تعیین گردیده و مطابق آن، ویژگی‌های کلی پیکره، نوع داده‌ها و فراداده‌ها، معیارهای مرتبط با فارسی‌آموز و نوع نشانه‌گذاری مشخص شده است. جدول ۱ برخی از معیارهای طراحی پیکره سلام فارسی را نشان می‌دهد.

جدول ۱. برخی از معیارهای طراحی پیکرهٔ «سلام فارسی»

۱	رسانه:	نوشتاری	۶	نوع برچسب:	خطاهای زبانی
۲	حجم:	۲۷ هزار واژه	۷	نوع داده:	متن نگارشی
۳	کاربرد:	پژوهشی	۸	نوع تکلیف:	انشاء، متن آزاد
۴	دسترسی	برخط	۹	ژانر متن:	توصیفی، روایی
۵	زبان اول	صربی	۱۰	سطوح زبانی	پایه تا پیشرفته (۴ سطح)

پس از تعیین معیارهای طراحی پیکره که بر مبنای الگوی پیشنهادی طراحی پیکره‌های زبان‌آموز (صفری، ۱۳۹۵) [۵] صورت گرفته است، فراداده‌های مرتبط با زبان‌آموز (از جمله سن، جنسیت، ملیت، سطح تحصیلات، مدت زمان فارسی‌آموزی و غیره) و فراداده‌های مربوط به متن (از جمله عنوان متن، سال تولید، شهر و کشوری که متن تولید شده و غیره) نیز

1 Faculty of Philology, University of Belgrade

جمع‌آوری و نشانه‌گذاری گردید. نرم افزار پیکره داری چهار ابزار برای ارسال و ثبت متن‌های خام در پایگاه داده‌ها، برچسب‌زنی فراداده‌ها و خطاها، ابزار فیلتر/ جستجو و ابزار نمایش آمار و گزارش‌های پیکره تشکیل شده و دسترسی به این ابزارها به صورت برخط در وبگاه پیکره[1] ممکن است. متن‌های پیکره پس از ترانویسی و تبدیل به دادهٔ الکترونیکی قابل خوانش و نشانه‌گذاری، در پایگاه داده‌های پیکره[2] ذخیره می‌شوند و سپس با استفاده از ابزار برچسب‌زنی خطاها، امکان برچسب‌دهی بر روی متن فراهم می‌آید. هر متن پیکره به عنوان یک سند مجزا با فراداده‌هایی که مشخصات متن و تولید کننده (فارسی‌آموز) را دارد در پیکره ثبت می‌شود و مجموعهٔ متن و برچسب‌ها به تفکیک برای هر سند و همچنین تمامی اسناد در دو قالب پی دی اف و متن خام، قابل بارگیری هستند (نمونه در پیوست ۱). تاجاییکه نگارنده مطلع است، این پیکره نخستین پیکره مدون زبان‌آموز فارسی است که با هدف نشانه‌گذاری و تعیین خطاهای فارسی‌آموزی طراحی و تولید شده است.

۳. الگوی دسته‌بندی و تعیین خطاهای پیکره

به کارگیری پیکره‌های زبان‌آموز برای شناسایی و بررسی خطاهای زبانی، در حوزه مطالعات نظری تحلیل خطا یا تحلیل مقابله‌ای خطاها قرار می‌گیرد، با این حال، از آنجاییکه این نوع از پیکره‌های زبانی به شیوهٔ نظام‌مند داده‌های خام زبانی را به شکل الکترونیکی ذخیره می‌کنند که قابل نشانه‌گذاری و خوانش با استفاده از نرم‌افزارهای پیکره هستند، به رویکرد «تحلیل خطا با کمک رایانه»[3] نیز مرتبط می‌شوند. بر همین اساس، گرانجر (۲۰۰۲) [۶] تحلیل مقابله‌ای زبان میانی و تحلیل خطا با کمک رایانه را «روش شناسی» قدرتمند برای مطالعه کمی و کیفی زبانِ زبان‌آموز می‌داند زیرا چنانکه اشاره شد، در این شیوه از ابزارها و نرم‌افزارهای رایانه‌ای برای ذخیره و پردازش خطاها استفاده می‌شود و در نتیجه، تحلیل‌های مبتنی بر آن عینی و دقیق‌تر خواهد بود.

از آنجایی که خطاهای زبانی را می‌توان از جنبه‌های گوناگون و با معیارهای متفاوت دسته‌بندی کرد، مهمترین موضوع پس از جمع‌آوری داده‌ها و تشکیل پیکرهٔ زبان‌آموز، تعیین

1 http://sflc.fil.bg.ac.rs
2 PostgresSQL
3 computer-aided error analysis

الگوی نظری برای دسته‌بندی خطاهای زبانی، تهیه مجموعه برچسب و ایجاد ابزار ویرایشگر خطا است تا به کمک آنها پیکره نشانه‌گذاری شود. تعیین الگوی دسته‌بندی خطاها از این جهت اهمیت دارد که مبنای تهیه مجموعه برچسب‌ها برای نشانه‌گذاری پیکره است و در نتیجه گزارش و تحلیل‌های پیکره عیناً بر اساس این الگو شکل می‌گیرد.

برای دسته‌بندی خطاهای زبانی، الگوهای متعددی معرفی شده است که هر کدام با توجه به اهداف پیکره می‌تواند مورد استفاده قرار گیرد، برای مثال لی (۱۹۹۰) [۷] الگوی چهار سطحی از خطاها را در دسته‌بندی‌های: (۱) خطاهای دستوری، (۲) خطاهای گفتمانی، (۳) خطاهای آوایی و (۴) خطاهای واژگانی، پیشنهاد می‌کند و ساویلی-ترویکی (۲۰۰۶) [۸] الگوی سه سطحی شامل (۱) خطاهای زبانی (آوایی، صرفی، نحوی، ..)، (۲) خطاهای عمومی در ساخت زبان (ساخت منفی، ساخت شرطی و..)، (۳) خطا در عناصر خاص زبانی (حروف اضافه، تعریف، صورت فعلی و..) را مناسب می‌داند. از آنجاییکه الگوی خطایابی پیکره برای زبان‌فارسی پیشینه‌ای ندارد، در پیکرهٔ سلام فارسی با استناد به الگوی «دسته‌بندی توصیفی خطاها» (دالی، برت و کرشن، ۱۹۸۲) [۹]، و با بهینه‌سازی آن برای زبان‌فارسی، دسته‌بندی سه سطحی برای تعیین خطاهای فارسی‌آموزان در پیکره انتخاب گردید که عبارتست از: (۱) خطاهای رو ساخت (حذف، اضافه، جابجایی، جایگزینی)، (۲) حوزهٔ خطا (نگارشی، صرفی، نحوی، واژگانی، سبکی) و (۳) نوع خطا (شامل ۲۲ نوع خطای مرتبط با ساخت فارسی). جدول ۲ الگوی دسته‌بندی خطاها در پیکره سلام فارسی را نشان می‌دهد.

جدول ۲. الگوی دسته‌بندی خطاها در پیکرهٔ «سلام فارسی»

Errors in Surface Structure	Addition, Omission, Substitution, Permutation
Error Domains	Orthography, Morphology, Syntax, Lexis, Style
Error Types	Consonant Character(s), Long Vowel Character(s), Short Vowel character(s), Connections, the Ezâfe Particle, Dots, Adjective, Noun-Plural, Noun (other), Pronoun, Preposition, Postposition (râ), Conjunction, Verb Tense, Verb Agreement, Verb (other), Adverb, Word Order, Word Selection, Phrase Selection, Cohesion and Unclear Style.

۴. مجموعه برچسب‌های پیکره

مجموعه برچسب‌های پیکره بر اساس الگوی دسته‌بندی خطاها (جدول ۲) و در مجموع با تعیین ۳۱ برچسب تعیین گردید تا خطاها در سه سطح و بر اساس الگوی تعیین خطاها نشانه‌گذاری شوند. هر خطا برچسبی ترکیبی از ۴ نشانه/حرف را شامل می‌شود. حرف اول، خطا، بیانگر عنوان خطا در سطح روساخت، حرف دوم، نشانهٔ حوزه خطا و دو حرف پایانی، بیانگر نوع خاص خطا هستند. قابلیت انتخاب و ترکیب خطاها در سه سطح به صورت آزاد و منعطف در مجموعه برچسب‌ها پیش‌بینی شده است و این امکان وجود دارد که خطاها در لایه‌های مختلف انتخاب و ترکیب شوند. برای مثال برچسب خطای <O_M_VT> بیانگر آن است که در لایه اول، خطای روساخت مربوط به «حذف» یک عنصر زبانی، در لایه دوم، حوزه خطا مربوط به «صرف» و در لایه سوم، نوع خطا «زمان فعل» است که در نهایت مشخص می‌کند، خطای نشانه‌گذاری شده مربوط به حذف یک عنصر زبانی در صرف فعل و مرتبط با تطابق زمانی فعل است و به عبارتی، در این خطا شناسه زمانی فعل حذف شده است. جدول ۳ فهرست مجموعه برچسب‌های پیکرهٔ سلام فارسی را مشخص می‌کند.

جدول ۳. مجموعه برچسب‌های پیکرهٔ سلام فارسی

First Level		Second Level		Third Level	
Surface Structure	Abbr	Error Domain	Abbr	Error Type	Abbr
Addition	A	Orthography	O	Consonant character(s)	CC
Omission	O	Morphology	M	Long Vowel character(s)	VL
Substitution	S	Syntax	S	Short Vowel character(s)	VS
Permutation	P	Lexis	L	Connections	CO
		Style	T	Ezâfe Particle	EP
				Dots	DT
				Adjective	AJ
				Noun-Plural	NP
				Noun Other	NO
				Pronoun	PR
				Preposition	PP
				Postposition (râ)	PO
				Conjunction	CN
				Verb Agreement	VA
				Verb Tense	VT
				Verb Other	VO
				Adverb	AD
				Word Order	WO
				Word Selection	WS
				Phrase Selection	PS
				Cohesion	CS
				Unclear style	US

۶- نمونه‌های زیر چگونگی خوانش خطاهای ثبت شده در پیکره هستند:

شناسه متن و زبان‌آموز: LIN022_T0088

*بلگراد از <A_S_PP > دو میلیون جمعیت دارد

برچسب خطاها: <A_S_PP >

Addition_Syntax_Preposition

خطای روساخت: حذف/ حوز خطا: نحو/ نوع خطا: حرف اضافه

شناسه متن و زبان‌آموز: LIN054_T0151

*به آرایشگاه{ آرایشگاه <S_O_CC > }رفتم

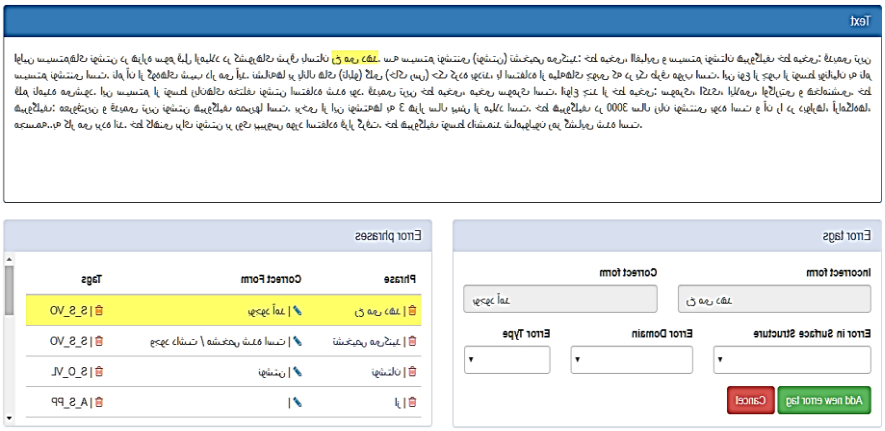
برچسب خطاها: < S_O_CC >

Substitution_Orthography_Consonant character(s)

خطای روساخت: جایگزینی/ حوزه خطا: نگارش/ نوع خطا: همخوان

۵. ابزار برچسب‌زنی خطاها

یکی از ابزارهای مورد استفاده در پیکره، «ابزار برچسب زنی خطا»[1] است که به منظور تسهیل در برچسب‌دهی خطاهای پیکره طراحی شده است. این ابزار مطابق با مجموعه برچسب‌های پیکره (جدول ۲) تهیه شده است و عملکردهای آن عبارتند از: (۱) انتخاب واژه/ها، عبارت/ها و جمله/ها برای نشانه‌گذاری خطاها، (۲) پیشنهاد صورت صحیح و ثبت در پیکره، (۳) انتخاب خطا و برچسب دهی در سه لایه مطابق مجموعه برچسب و (۴) ویرایش یا حذف خطای ثبت شده. ساختار ابزار برچسب زنی خطاها از سه بخش «جعبه متن»، «جعبه برچسب خطاها» و «جعبه نمایش گروه خطاها و برچسب‌ها» تشکیل شده با این حال به صورت یکپارچه عمل می‌کند و پس از برچسب‌دهی و تایید کل متن، داده‌ها را در پایگاه داده‌ها ثبت و ذخیره می‌کند. شکل ۱ نمایی از ابزار برچسب‌زنی و ثبت داده‌های نشانه‌گذاری شده در پیکره را نشان می‌دهد.



شکل ۱. نمای ابزار برچسب زنی پیکره

1 The Error Tagging Tool (ETT)

۶. فراوانی و نوع خطاها

پیکرهٔ سلام فارسی با هدف تعیین نوع و فراوانی خطاهای زبانی تشکیل شده است و چنانکه در بخش ۲ مقاله اشاره گردید، در فاز اول این پروژه، تولیدات نوشتاری فارسی‌آموزان صرب، جمع‌آوری و نشانه‌گذاری گردیده است. ابزار «آمار و گزارش‌دهی پیکره»[1] امکان ارایهٔ آمارهای تفکیکی از فراونی، نوع خطا را با ترکیب فیلترهای گوناگون از جمله سطح مهارت زبانی، جنسیت، سن، نوع متن و .. فراهم می‌سازد. به طور اجمالی، گزارش فراوانی و نوع خطاهای برچسب خورده در این پیکره به شرح زیر است:

۱. بیشترین خطا مربوط در «رو ساخت» مربوط به خطای جایگزینی (۴۵٪ مجموع خطاها) است.

۲. بیشترین خطاهای مربوط به «حوزه خطا» مربوط به حوزهٔ نگارش (۳۱٪) و سپس خطای نحوی (۲۸٪) است.

۳. بیشترین خطاهای مربوط به «نوع خطا»، مربوط به همخوان‌ها/ واکه بلند، ترتیب واژگانی و فعل هستند.

با استناد به گزارش‌های پیکره، می‌توان به این نتیجه رسید که خطاهای فارسی‌آموزان صرب در یادگیری زبان فارسی در دو بخش کلی قابل دسته‌بندی است:

الف) خطاهای نگارشی (املاء فارسی): جایگزینی نادرست همخوان‌های فارسی بیشترین خطای ثبت شده در حوزهٔ خطاهای نگارشی است. این خطاها عمدتا مربوط به املاء واژه‌هایی است که در آن واج مورد نظر دارای نویسه‌های چندگانه (مانند واج های/z/، /s/ و..) است و فارسی‌آموزان نویسه‌ها ناصحیح را جایگزین کرده‌اند.

ب) خطاهای ساخت زبانی: در این بخش، ترتیب نادرست واژگانی (به ویژه ترکیب‌های واژگانی با نقش نمای اضافه)، ساخت فعل‌های مرکب، واژه‌بست نشانهٔ نکره (ی) و کاربرد حروف اضافه، بیشترین خطاهای فارسی‌آموزان در کل پیکره را تشکیل می‌دهند.

در مجموع، خطاهای فارسی‌آموزان در سطوح مختلف زبانی متفاوت است و الگوی قابل تشخیص از گزارش‌های پیکره نشان می‌دهد که با افزایش سطح مهارت زبانی، تعداد کمّی خطاها کاهش می‌یابد و در سطوح بالاتر، عمدهٔ خطاها مربوط به حوزه نحو زبان فارسی ثبت

1 The Data Statistics Tool (DST)

شده که دلیل آن عدم تسلط فارسی‌آموزان در استفاده از ساخت‌های پیچیده‌تر دستوری است.

۷. نتیجه‌گیری

پیکره‌های زبان‌آموز بر اساس معیارهای مشخصی طراحی و تولید می‌شوند، یکی از معیارهای ساختاری پیکره‌ها، نوع نشانه‌گذاری پیکره است. در پیکره‌های خطایابی که عمدتاً با دو هدف تعیین فراوانی و تشخیص نوع خطاهای زبان‌آموزان تشکیل می‌شود، تهیه «الگوی دسته‌بندی خطاها» و «ایجاد مجموعه برچسب خطاها»، دو رکن اصلی پیکره خطایابی زبان‌آموز را تشکیل می‌دهند زیرا دریافت گزارش‌ها و تحلیل‌های مبتنی بر پیکره، بطور مستقیم به این دو مهم وابسته است. در پیکرهٔ «سلام فارسی»، پس از تعیین این دو رکن، با استفاده از ابزار تسهیل‌کننده برچسب‌دهی، نشانه‌گذاری خطاهای فارسی‌آموزان در سه لایه مستقل و در عین حال پیوسته صورت می‌گیرد و بدین ترتیب هر برچسب، ویژگی‌های خطا در روساخت، حوزه و نوع آن را مشخص می‌کند. با استناد به گزارش‌های پیکره‌های خطایابی زبان‌آموز می‌توان شناخت بهتری از دشواری‌های یادگیری زبان فارسی برای گروه‌های گوناگون فارسی‌آموزان به‌دست‌آورد و با مطالعه درزمانی و بررسی خطاها در سطوح متفاوت مهارت زبانی، از نیازهای آموزشی ایشان مطلع شد. استفاده از گزارش‌ها و تحلیل‌های مبتنی بر پیکره‌های خطایابی زبان‌آموز علاوه بر آنکه داده‌ها و منابع پژوهشی جدید برای مطالعات نظری مربوط به فراگیری زبان دوم را فراهم می‌سازد، در بعد عملی و کاربردی نیز امکان برنامه‌ریزی هدفمند و تعیین برنامه‌درسی مناسب، تعیین راهبردهای و شیوه‌های تدریس و نهایتاً تهیه و تولید محتوای آموزشی مناسب برای گروه‌های گوناگون زبان‌آموزان را فراهم می‌سازد.

قدردانی

پیکرهٔ زبان‌آموز «سلام فارسی» با حمایت دانشکدهٔ ادبیات و زبان‌های خارجی (فیلولوژی) دانشگاه بلگراد و نظارت علمی خانم دکتر مایا میلیچوویچ[1]، دانشیار رشته زبان‌شناسی و متخصص حوزهٔ زبان‌شناسی پیکره‌ای، طراحی و تشکیل شده است؛ همچنین در تحلیل و بررسی گزارش‌های پیکره از راهنمایی آقای دکتر رضامراد صحرایی، دانشیار رشته زبان‌شناسی

1 Maja Miličević

و آموزش زبان فارسی به غیرفارسی زبانان دانشگاه علامه طباطبایی، بهره‌مند شـده‌ام و زحمـات این استادان را ارج می‌نهم.

منابع

[39]Granger, S. (2008). Learner Corpora. In A. Ludeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (pp. 259–275). Berlin, Germany: Walter de Gruyter.

[40]Callies, M., Paquot, M. 2015." Learner Corpus Research: An interdisciplinary field on the move". *International Journal of Learner Corpus Research*, 1, pp.1-6.

[41]Granger, S., Gilquin, G., & Meunier, F. 2015. "Learner Corpus Research Past, Present and Future". In book: S. Granger, G. Gilquin & F. Meunier, *The Cambridge Handbook of Learner Corpus Research* Cambridge University Press: Cambridge. pp1-5.

[42]Nesselhauf N. (2004a), Learner corpora and their potential in language teaching. In: J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (125-152). Amsterdam: Benjamins.

[۴۳]صفری، سعید، ۱۳۹۵. "پیکرهٔ زبان‌آموز، مبانی، روش‌شناسـی، الگـوی طراحـی و تولیـد". *مجموعه مقـالات دومـین همـایش ملـی زبـان‌شناسـی پیکـره‌ای*، آزاده میرزایـی. تهـران. نشرنویسه‌پارسی. صص ۹۳-۱۲۴.

[44]Granger, S. 2002. "A bird's-eye view of computer learner corpus research". In: S. Granger, J. Hung, S. Petch-Tyson & J. Hulstijn (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam & Philadelphia: John Benjamins. pp. 3-33

[45]Lee, N. 1990. "Notions of Error and Appropriate corrective Treatment", *Hong Kong Papers in Linguistics and Language Teaching*, 14, pp.55-70.

[46]Saville-Troike, M. 2006. *Introducing Second Language Acquisition*. Cambridge. Cambridge University Press.

[47]Dulay, H. C., Burt, M. K. & Kreshen, S. 1982. *Language Two*. New York: Oxford University Press.pp.150-160

پیوست ۱

نمونه‌ای از متن فارسی‌آموز و ثبت خطاها در پیکرهٔ سلام فارسی. (بارگیری سند در قالب متن خام از پایگاه ذخیره داده‌های پیکره).

LIN005_T0010_NA_C1_sr_RS

Text:

اسب سیاهی در چمن زارم می چرد.
در شب من به چشمن مثل دزد باورچین باورچین نزدیک می شوم.
می بینم:
اسب سفیدی با خورشید در پیسانی و بالهای نقره آبی در چمن می ایستد.
من بر بلهایش به ابدیت خیالاتی می پرم.
او برای من شعر می سراید
ستاره من ارازی آن را می خواند.
در صبح زود اسب سیاهی در چمن زارم می چرد.

Errors:

13 | چمن زارم | چمن زار | A_M_PR
30 | در شب من به چشمن مثل دزد باورچین باورچین نزدیک می شوم | | S_T_CS
42 | چشمن | چشمان | O_O_VL
107 | خورشید | خورشید | O_O_DT
117 | پیسانی | پیشانی | O_O_DT
165 | بلهایش | بالهایش | O_O_VL
175 | ابدیت خیالاتی | خیالات ابدی | P_S_AJ
221 | ستاره من ارازی آن را می خواند | | S_T_US

بهبود برچسب‌گذاری اجزای کلام با استفاده از نرم‌افزار رفع ابهام‌کننده از برچسب هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»

الهام علایی ابوذر*

چکیده

در پژوهش حاضر به این مسئله پرداخته شد که آیا با رفع ابهام از برچسب نحوی هم-نگاره‌های اسمی و صفتی مختوم به «-ی»، که فراوانی بالایی در پیکره‌های متنی فارسی دارند، کارایی یک سیستم برچسب‌زنی خودکار اجزای کلام، افزایش می‌یابد؟ سیستم برچسب‌زنی مورد مطالعه در پژوهش حاضر، ابزار «هضم» است. در پژوهش حاضر ابتدا فهرست مبسوطی از هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» با تعریف تعداد ۱۰ پنجره، به عبارتی دیگر، ۱۰ کلمه قبل و بعد از هر هم‌نگارهٔ مختوم به «ی»، در پیکرهٔ بی‌جن‌خان (که پیکره‌ای است با برچسب اجزای کلام) تهیه شد؛ پس از بررسی هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» در بافت نحوی، الگوهای حساس به بافت نحوی جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های مذکور، استخراج شد؛ سپس، نرم‌افزاری جهت رفع ابهام از برچسب نحوی این هم‌نگاره‌ها، تهیه شد. ارزیابی کلی نرم‌افزار تهیه شده جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» در فارسی، نشان می‌دهد اگر تنها الگوهای حساس به بافت نحوی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به برچسب‌زن «هضم» اضافه کنیم، صحت (Accuracy) کلی برچسب‌زن ۹۵,۹۶۱ درصد می‌شود که ۱,۳۴ درصد نسبت به حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، بالاتر است.

واژه‌های کلیدی: برچسب نحوی، هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»، الگوهای حساس به بافت نحوی، صحت برچسب‌زنی

* عضو هیات علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) Alayi@irandoc.ac.ir

۱- مقدمه

در زبان‌شناسی رایانشی، برچسب‌گذاری اجزای کلام، درواقع عمل انتساب برچسب به کلمات تشکیل‌دهندهٔ یک متن یا یک پیکره است. این برچسب‌گذاری براساس مقولهٔ آن کلمه در متن (مانند « اسم»، «فعل»، «قید»، «صفت»، و غیره) صورت می‌پذیرد. برچسب‌گذاری اجزای واژگانی کلام، از پیش‌نیازهای بسیاری از فعالیت‌های حوزهٔ پردازش زبان طبیعی[1] از جمله ترجمهٔ ماشینی، خطایابی، تبدیل متن به گفتار، بازیابی اطلاعات و کمک به مدل‌های آماری است ([1] مگردومیان: ۲۰۰۴). همچنین در طراحی سیستم نمایه‌ساز ماشینی، یکی از بخش‌ها، طراحی زیرسیستم تحلیل واژگانی است. این زیرسیستم، متن را به واژه‌ها تفکیک می‌کند و ماهیت هر کلمه را تشخیص می‌دهد و تشخیص نوع واژه و شناسایی فعل‌ها، الفاظ و اصطلاح‌ها را در بر دارد. بنابراین سیستم برچسب‌دهی خودکار، ماهیت مقولهٔ کلمات را مشخص می‌کند تا بتوان در مراحل بعدی از این اطلاعات در جهت استخراج کلیدواژه‌ها یا هر نوع بازیابی اطلاعات از متن، استفاده کرد. سامانه‌های برچسب‌گذاری، به دلیل عدم اشراف کامل به قواعد ساخت‌واژی زبان، ممکن است در برخورد با کلمات دارای پیچیدگی‌های ساخت‌واژی، با مشکلاتی مواجه شوند ([2] محسنی: ۱۳۸۷). زبان فارسی نیز دارای پیچیدگی‌هایی است که مشکلاتی بسیاری را در مسیر برچسب‌گذاری رایانه‌ای اجزای واژگانی کلام ایجاد می‌کند. یکی از این پیچیدگی‌ها مربوط به شکل یکسان برخی از تکواژها است که باعث ابهام در متون فارسی می‌شود. بعضی کلمات در پیکره‌های متنی ممکن است بیش از یک برچسب داشته باشند؛ زیرا کلمات در جایگاه‌های مختلف می‌توانند برچسب‌های واژگانی متفاوت داشته باشند. مانند موارد زیر:

الف. برچسب یای نکره : کشاورزی را دیدم.

ب. برچسب یای اسم‌ساز: وی چند سال کشاورزی را دنبال می‌کرد.

پ. برچسب یای شناسه دوم شخص مفرد: تو خود کشاورزی.

همچنین در فارسی هم‌نگاره‌های[2] بسیاری به دلیل پیچیدگی‌های موجود در ساخت‌واژهٔ

---

1 Natural language processing
2 homographs

فارسی، به وجود می‌آیند. هم‌نگاره‌ها کلماتی هستند که صورت نوشتاری یکسان، اما منشاء، معنی یا تلفظ متفاوت دارند ([3] فرهنگ لغت مریم وبستر[1]). در زبان‌هایی که ساخت‌واژهٔ پیچیده دارند، مانند زبان فارسی، هم‌نگاره‌های بسیاری ساخته می‌شوند. گویشوران فارسی به دلیل مجهز بودن به اطلاعات زبانی از قبیل اطلاعات مربوط به ساخت‌واژهٔ فارسی، ساخت واجی فارسی و ساخت نحوی فارسی، قادر به رفع ابهام از هم‌نگاره‌ها در بافت نحوی هستند، اما سامانه‌های پردازش متون فارسی، به دلیل عدم دسترسی کامل به چنین اطلاعات زبانی، با مشکلاتی در پردازش هم‌نگاره‌ها مواجه می‌شوند که یکی از این مشکلات اختصاص برچسب اجزای کلام درست به هم‌نگاره‌ها در متون فارسی است. بررسی کلی هم‌نگاره‌ها در پیکره‌های متنی موجود فارسی نشان می‌دهد که تعداد هم‌نگاره‌ها در پیکره‌ها قابل توجه است. اکثر این هم‌نگاره‌ها، در اثر یکسان بودن نمود نوشتاری تکواژ یای نکره، یای اسم‌ساز (اسم مکان، اسمی که دال بر شغل یا محافظت و دارندگی است، اسم‌معنی یا اشیا، تصغیر و تحبیب، اسم مصدر یا حاصل مصدر)، شناسهٔ دوم شخص مفرد و یای صفت‌ساز (صفت فاعلی و مفعولی، صفتی که دال بر نسبت است) و یای متصل به گروه اسمی به وجود آمده‌اند ([4] علایی : ۱۳۹۵). سوال مطرح در پژوهش حاضر این است که آیا با رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»، که فراوانی بالایی در پیکره‌های متنی فارسی دارند، کارایی یک سیستم برچسب‌زنی خودکار (مطالعهٔ موردی: سیستم برچسب‌دهی خودکار «هضم»)، افزایش می‌یابد؟ «هضم» ابزاری است جهت پردازش زبان فارسی با استفاده از زبان برنامه‌نویسی پایتون[2] که برای پیش‌پردازش‌هایی چون نرمالایزکردن متن، تقطیع جمله‌ها و واژه‌ها، ریشه‌یابی واژه‌ها، تحلیل صرفی واژه‌ها و تجزیهٔ نحوی جمله مورد استفاده قرار می‌گیرد. (sobhe/hazm[5]).

۲- روش انجام پژوهش

به منظور دستیابی به هدف پژوهش، یعنی «بررسی امکان بهبود بخشیدن به یکی از سیستم‌های موجود برچسب‌دهی اجزای کلام، به نام «هضم»، از طریق به کار بردن نرم‌افزار

1 Merriam Webster
2 Python

مجهز به رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم بـه «ی» در فارسـی»، ارزیابی های متنوعی انجام شده است که مبنای آنها دو معیـار کلـی «صـحت[1]» و «معیـار اِف[2]» است. برای پیاده‌سازی قوانین از زبان برنامه‌نویسی پایتون اسـتفاده شـده اسـت. بـا جسـتجو در برچسب‌های اجزای کلام اعمال‌شده به مـتن ورودی توسـط برچسـب‌زن هضـم، عبـارات دارای الگوهای مدنظر مشخص شد و وضعیت انطباق آن‌ها با قوانین ارائه‌شده مورد بررسی قرار گرفت. برنامه به این صورت عمل می‌کند که با پیمایش درکلیهٔ واژه‌های متن، هرگـاه بـا واژه‌ای مواجـه شود که به «ی» ختم می‌شود، تک‌تک الگوها را در مورد این واژه بررسی می‌کند. به ایـن معنـا که این واژه را به عنوان واژهٔ مختوم به «ی» در هر الگو در نظر می‌گیرد و بررسی می‌کند که آیا با الگوی ارائه‌شده همخوانی دارد یا خیر. به منظور آشنایی با نحـوهٔ عملکـرد برنامـه، چگـونگی پیاده‌سازی دو مورد از الگوها برای نمونه تشریح می‌شود:

۱- زبان/ لهجه+ صفت(ADJ)

مثال از پیکرهٔ متنی:

....آشنایی کامل به چند زبان بین‌المللی (ADJ) از امتیازهای لازم...

۲- فعل (V) + حرف عطف (CON) / ویرگول (،) (DELM) +***اسم (N)*** + حرف ربط «که»

مثال از پیکرهٔ متنی:

.....انجام می‌داد (V) ، (DELM) دوره‌ای (N) که تهران اقتصاد روستایی داشت....

برای یافتن نمونه‌های منطبق با این الگوها بررسی می‌شود کـه مـثلاً در الگـوی ۱، آیـا واژهٔ قبل از واژهٔ مختوم به «-ی»، واژهٔ «زبان» یا «لهجه» است؟ اگر اینگونه بود برچسب «صفت» به این واژه اختصاص داده می‌شود. به منظور ارزیابی دقیق تأثیر افزودن الگوهای استخراج‌شده بـه برچسب‌زن اجزای کلام، آزمایش‌های مختلفی بر اساس دو معیار «صحت» و معیار «اف» انجـام شد؛ به ساده‌ترین شکل، معیار «صحت» را می‌توان اینگونه تعریف کرد: از مجموع برچسب‌هایی که سیستم به واژه ها (و علائم نگارشی) اختصاص داده است، چند درصد صحیح هسـتند یـا بـه

1 Accuracy
2 F-Measure

عبارتی چند درصد با برچسب‌های درست، که توسط نیروی انسانی مشخص شده‌اند، انطباق دارند. و معیار «اف» نیز از ترکیب دو معیار دقت[1] که با فرمول زیر محاسبه می‌شود

$$\text{دقت (P)} = \frac{\text{تعداد برچسب‌های x که سیستم به درستی تشخیص داده است}}{\text{تعداد کل برچسب‌های x تشخیص داده شده توسط سیستم}}$$

و فراخوانی[2] که با فرمول زیر محاسبه می‌شود

$$\text{فراخوانی (R)} = \frac{\text{تعداد برچسب‌های x که سیستم به درستی تشخیص داده است}}{\text{تعداد کل برچسب‌های x در داده‌ای که نیروی انسانی برچسب زده است}}$$

به دست می‌آید. بنابراین براساس دو معیار فوق، معیار «اف» به شکل زیر محاسبه می‌شود:

$$\text{معیار اف} = \frac{2 \times P \times R}{P+R}$$

با توجه به اینکه قواعد مورد نظر بر اساس پیکرهٔ بی‌جن‌خان استخراج شده بود، برچسب‌زن هضم با همین پیکره آموزش داده شد. برای این منظور برچسب‌زن با دادهٔ آموزش (۰,۹ بی‌جن‌خان) آموزش داده شده و برچسب‌زنی بر روی دادهٔ آزمون (۰,۱ بی‌جن‌خان) انجام و برچسب‌های هضم با برچسب‌های انسانی دادهٔ آزمون مقایسه شده‌اند. ([6] در جورافسکی و مارتین ۲۰۰۹، توضیحات مربوط به معیار صحت (Accuracy) و داده‌های آزمون و آموزش در فصل ۵، و توضیحات مربوط به معیار« اف» در فصل ۱۳ ارائه شده است). به منظور ارزیابی دقت برچسب‌دهی با در نظر گرفتن الگوهای رفع ابهام از هم‌نگاره‌های اسمی و صفتی مختوم به «ی»، متنی که پیش از این به ابزار هضم به عنوان ورودی وارد شده بود و برچسب‌گذاری شد، این بار با در نظر گرفتن الگوها مجدداً برچسب‌ها مورد بازبینی قرار گرفت. در این آزمایش دقت برچسب‌زن اجزای کلام هضم به صورت کلی (در سطح تمامی برچسب‌ها) با استفاده از معیار صحت و در دو حالت «با الگوها» و «بدون الگوها» انجام شده است. در حالت «با الگوها» تمامی

1 Precision
2 Recall

الگوهای استخراج‌شده در طرح پژوهشی «*رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی* ([7] علایی: ۱۳۹۵) » در هنگام برچسب‌زنی به کار گرفته شدند. جدول ۱، نتیجه این بررسی را نشان می‌دهد:

جدول ۱: درصد صحت (Accuracy) کلی برچسب‌زن هضم

بدون الگوها	با الگوها
۹۵,۶۹۰	۹۴,۳۵۱

همانطور که ملاحظه می‌شود به‌کارگیری همهٔ الگوهای استخراج‌شده موجب کاهش ۱,۳۳۹ درصدی صحت برچسب‌زنی شده است. در مرحله بعد، این بار تأثیر هر قاعده/ الگو در بهبود دقت برچسب‌زن مورد بررسی قرار گرفت؛ هدف از این آزمایش بررسی تأثیر تک‌تک الگوها بر بهبود دقت برچسب‌زن بوده است. برای محاسبهٔ تأثیر هر قاعده، معیار صحت بر اساس تمام مواردی که قاعده اعمال شده است، یک بار بدون الگوها و یک بار با الگوها محاسبه شد. در این مرحله مشخص شد که برخی از الگوها دقت برچسب‌دهی را کمی بالا می‌برند. قسمتی از نتیجهٔ این بخش در جدول ۲ آورده شده است:

جدول ۲: میزان تأثیر تک‌تک الگوها

قاعده	بدون الگوها	با الگوها	میزان تاثیر
rule01a	۰.۹۷۶۵۹	۰.۸۷۹۹۲	-۰.۰۹۶۶۸
rule01b	۰.۹۷۰۷۶	۰.۸۳۰۴۱	-۰.۱۴۰۳۵
rule02a	۰.۸۷۵۰۰	۰.۶۳۳۳۳	-۰.۲۴۱۶۷
rule02b	۱.۰۰۰۰۰	۰.۸۳۳۳۳	-۰.۱۶۶۶۷
rule03	۰.۹۰۰۰۰	۰.۴۲۵۰۰	-۰.۴۷۵۰۰
rule05a	۰.۸۳۷۸۴	۰.۴۰۵۴۱	-۰.۴۳۲۴۳
rule06a	۱.۰۰۰۰۰	۱.۰۰۰۰۰	۰.۰۰۰۰۰
rule06b	۰.۷۵۰۰۰	۱.۰۰۰۰۰	+۰.۲۵۰۰۰
rule07	۰.۸۸۲۳۵	۰.۴۳۵۲۹	-۰.۴۴۷۰۶
rule08	۰.۹۳۹۳۹	۱.۰۰۰۰۰	+۰.۰۶۰۶۰
rule08DELM_hazm	۰.۹۵۷۴۵	۰.۷۲۳۴۰	-۰.۲۳۴۰۴
rule09a	۰.۹۴۶۵۰	۰.۹۳۶۳۱	-۰.۰۱۰۱۹
rule09b	۰.۸۷۱۴۶	۰.۷۶۱۱۳	-۰.۱۱۰۳۲

لازم به ذکر است به علت کمبود فضا، همهٔ ۳۶ الگو در جدول ۲ آورده نشده است و تنها چند الگو به عنوان نمونه آورده شده است.

نهایتاً این نتیجه به دست آمد که اگر تنها الگوهایی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به برچسب‌زن اضافه کنیم صحت (Accuracy) کلی برچسب‌زن ۹۵,۹۶۱ درصد می‌شود که ۱,۳۴ درصد نسبت به حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، بالاتر است.

۳- نتیجه‌گیری

در پژوهش حاضر، ابتدا هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» در فارسی، در بافت نحوی مورد مطالعه قرار گرفت و الگوهای حساس به بافت نحوی جهت تخصیص برچسب نحوی صحیح به این هم‌نگاره‌ها، استخراج شد. سپس نرم‌افزاری تهیه شد که کاربرد الگوهای مذکور را با استفاده از معیارهای «صحت» و «اف» را در سیستم برچسب‌گذاری خودکار «هضم» می-سنجید، در این پژوهش دقت برچسب‌زن اجزای کلام هضم به صورت کلی (در سطح تمامی برچسب‌ها) با استفاده از معیار صحت و در دو حالت «با الگوها» و «بدون الگوها» انجام شد. نهایتاً این نتیجه به دست آمد که اگر تنها الگوهایی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به برچسب‌زن اضافه کنیم صحت (Accuracy) کلی برچسب‌زن ۹۵,۹۶۱ درصد می‌شود که ۱,۳۴ درصد نسبت به حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، بالاتر است.

منابع

[48]Megerdoomian, Karine 2004. Developing a Persian part-of-speech tagger. *Proceedings of the 1st workshop on Persian language and computer*. Pp. 99-105.

[۴۹]محسنی، مهدی (۱۳۸۷). *سیستم برچسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکره متنی زبان فارسی*. دانشگاه علم و صنعت. دانشکده مهندسی کامپیوتر

[۵۰] http://www.merriam-webster.com

[۵۱]علایی، ا. ۱۳۹۵. *بررسی ساخت‌واژی هم‌نگاره‌های اسمی و صفتی به منظور کمک به*

برچسب‌دهی «اسم» به کلیدواژه‌ها در پیکره‌های علمی. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).

[52]https://github.com/sobhe/hazm

[53]Jurafsky, D., & Martin, J. H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed., Prentice Hall Series in Artificial Intelligence). Upper Saddle River, New Jersey: Pearson Education.

[۵۴]علایی، ا. ۱۳۹۵. *رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی*. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).

ساخت هستان‌شناسی مفاهیم آواشناسی درفارسی با استفاده از پروتژه

پیمان محمدی کرمانی*، بهاره پهلوان‌زاده و محمدهادی فلاحی*****

چکیده

امروزه نقش و اهمیت هستان‌شناسی به طور فزاینده‌ای در حال گسترش می‌باشد. هستان‌شناسی با دارا بودن یک ساختار مناسب این امکان را بوجود آورده است تا بتوان از آن در حوزه‌های گوناگونی نظیر وب معنایی، هوش مصنوعی، ماشین ترجمه، علم کتابداری و اطلاع رسانی، زبانشناسی رایانشی، پایگاه داده، مدل‌سازی اطلاعات، ادغام اطلاعات و بازیابی اطلاعات استفاده نمود. شناسایی واژه‌ها و مفاهیم آواشناسی و یافتن روابط میان آنها جهت پیاده‌سازی در قالب زبان‌های اُ.دبلیو.ال و اچ.تی.ام.ال از اهداف هستان‌شناسی آواشناسی بوده است. هستان‌شناسی آواشناسی با استفاده از ابزارهای مهندسی پروتژه پدید آمده است. هستان‌شناسی آواشناسی شامل ۳۷۷ واژه آواشناسی می‌باشد که این واژه‌ها از تالیفات حق‌شناس، ثمره، بی‌جن‌خان، مشکوةالدینی و کریستال استخراج گردیده است و بر اساس روابط سلسله مراتبی از بالا به پایین با تعریف ویژگی‌هایی از قبیل «نوعی از»، «جزئی از»، «عضوی از»، «مرتبط با» توسعه یافته است. در مقایسه انجام شده هستان‌شناسی آواشناسی با دو هستان‌شناسی فارس‌نت و هستان‌شناسی گلد مشخص گردید که هستان‌شناسی آواشناسی دارای مزیت‌های نسبی نسبت به دو هستان‌شناسی دیگر می‌باشد.

واژه‌های کلیدی: اصطلاحنامه، هستان‌شناسی، آواشناسی

۱- مقدمه

هستان‌شناسی از علوم ریشه‌دار و قدیمی محسوب می‌گردد که ریشه در علم فلسفه دارد.

---

* دانشجوی کارشناسی ارشد، زبان‌شناسی رایانه‌ای، مرکز منطقه‌ای اطلاع رسانی علوم و فناوری، peymanmkr@gmail.com

** استادیار، گروه پژوهشی طراحی و عملیات سیستم‌ها، مرکز منطقه‌ای اطلاع رسانی علوم و فناوری (نویسنده مسئول) pahlevanzadeh@ricest.ac.ir

*** استادیار، گروه زبان‌شناسی رایانه‌ای، مرکز منطقه‌ای اطلاع رسانی علوم و فناوری falahi@ricest.ac.ir

علم هستان‌شناسی در فلسفه به مطالعه چیستی جهان می‌پردازد و سعی در اکتشاف جهان پیرامون دارد اما معنای امروزی آن به کلی متحول گردیده است. امروزه هستان‌شناسی‌ها در پی یافتن روابط بین مفاهیم و اشیاء می‌باشد. در واقع هستان‌شناسی شامل واژگان نمایشی است که اغلب به برخی حوزه‌ها یا موضوعات خاص اختصاص پیدا می‌کند. هستان‌شناسی با یافتن روابط موجود بین واژه‌ها و مفاهیم سعی در آشکار کردن ماهیت هر چه بیشتر آنها می‌کند و به مثابه پیکره‌ای از واژگان و مفاهیم می‌باشد که حقایقی در مورد یک دامنه موضوعی خاص را مشخص می‌کند[۱].

هستان‌شناسی به عنوان یک زبان نمایش دانش عمل می‌کند که در آن دانش به اشتراک گذاشته می‌شود[۲]. در واقع هستان‌شناسی این امکان را بوجود آورده است که بتوان به پرس‌وجو در سطح دانش پرداخته و همچنین به تبادل اطلاعات بین عامل‌ها و سیستم‌های گوناگون اقدام نمود. قابلیت‌هایی نظیر بهبود سازماندهی دانش و دارا بودن یک ساختار مناسب امکان به کارگیری آن در وب معنایی، هوش مصنوعی، ماشین ترجمه، حوزه‌های تخصصی موضوعی، علم کتابداری و اطلاع رسانی، مهندسی دانش، نمایش دانش، زبانشناسی رایانشی، پایگاه داده، مدلسازی اطلاعات، ادغام اطلاعات، بازیابی و استخراج اطلاعات علوم دیگر فراهم آید.

اهداف ایجاد یک هستان‌شناسی از دیدگاه آسچولد[1] و کینگ[2] [۳] بسیار گسترده می‌باشد. اولین مورد از ضرورت ایجاد هستان‌شناسی این است که باعث می‌گردد انسان و ماشین از ساختار دانش، درک یکسان و مشترکی داشته باشند و در نتیجه این امکان به وجود آید که کامپیوتر و انسان تعامل بیشتری با یکدیگر داشته باشند. این تعامل به عنوان مثال در محیط وب به این طریق میسر می‌گردد که اطلاعات پراکنده موجود در مورد یک موضوع خاص توسط هستان‌شناسی تجمیع و بدین وسیله بازیابی اطلاعات آسان‌تر و کارآمدتر خواهد شد.

دومین هدف این است که با توجه به خاصیت بسط‌دهی و انعطاف‌پذیری هستان‌شناسی این امکان به وجود می‌آید که بتوان یک هستان‌شناسی را با توجه به نیاز خود باز طراحی نمود و حتی چندین هستان‌شناسی را در هم آمیخت و به این طریق دامنه دانش بروزتری را ارائه

1 Mike Uschold
2 Martin King

نمود.

هستان‌شناسی همچنین این مزیت را دارا می‌باشد که بتوان با استفاده از آنها بـه برداشـت صریح تری از یک دامنه خاص دست یافت و این به نوبه خود کمکی مضائفی خواهـد بـود بـرای نوآموزان و جستجوگران آن دانش خاص تا بینش بهتری نسبت به دانش مـورد مطالعـه داشـته باشند.

از سوی دیگر هستان‌شناسی موجب افزایش فهم مشترک در یک حوزه خاص می‌شـود. در واقع هستان شناسی به ما کمک می کند تا تفـاوت‌هـا، همپوشـانی و عـدم انطبـاق در مفـاهیم، ساختارها، اصطلاحات و غیره از میان برداشته شوند. به این ترتیب، هستان شناسی می تواند بـه عنوان یک چارچوب که دیدگاه‌های مختلف را متحد می کنـد و ارتباطـات را بهبـود مـی‌بخشـد مورد استفاده قرار گیرد.

به گفته چاندرسکاران[1]، جوزفسـون[2] و بنجـامینز[3] [۴] تجزیـه و تحلیـل هسـتان شناسـی ساختار دانش را شفاف می‌سازد. به گفته ایشان دامنه‌ای که بر اساس آن یـک هسـتان‌شناسـی شکل می‌گیرد قلب سیستم نمایش دانش برای آن دامنه را تشـکیل مـی‌دهـد. چاندراسـکاران و همکارانش معتقد به این امر می‌باشند که بدون وجود هستان‌شناسی یا مفهـوم‌سـازی‌هـایی کـه دانش را پایه گذاری می‌کنند، نمی توان واژگان پایه برای نمایش دانش ایجاد نمود. به نظـر آنهـا اولین گـام در طراحـی یـک سیسـتم نمـایش دانـش و واژگـان، انجـام تجزیـه و تحلیـل مـوثر هستان‌شناسی در آن زمینه یا دامنه است.

از دیگر مواردی که باعث گردیده است که هستان‌شناسی در دنیای مورد توجه قرار گیرنـد این است که با استفاده از هستان شناسی به اشتراک گذاشته شده، پایگاه های خاصی ایجاد شود تا اینکه موقعیت های خاصی را توصیف نمود. به عنوان مثال، رشته های مختلف دانش در ایـران می‌توانند از واژگان و نحو مشترک برای مفاهیمی که به دانش خاص اضـافه مـی‌گـردد اسـتفاده نمایند و سپس آنها را به اشتراک بگذارند. این نوع بـه اشـتراک گـذاری بـه طـور چشـمگیری توانایی استفاده مجدد دانش را افزایش می‌دهد.

با توجه به این امر که ایجاد هستان‌شناسی عمومی جامعی که تمام حـوزه‌هـا را دربرگیـرد

1 Balakrishnan Chandrasekaran
2 John R. Josephson
3 V. Richard Benjamins

امری مشکل و مستلزم به کارگیری گروه‌های تخصصی بی‌شماری می‌باشد و همچنین به دلیل ماهیت پیچیده زبان و اینکه هر مفهوم ممکن است دارای چندین وجه متمایز باشد و بنابراین ممکن است که هستان‌شناسی‌های عمومی نتوانند تمام موارد خواسته شده را به صورت مناسب پوشش دهند، این پژوهش دامنه موضوعی خود را محدود نموده و به طور تخصصی در حوزه هستان‌شناسی آواشناسی وارد گردیده است.

از آن جهت که این هستان‌شناسی در حوزه علم آواشناسی توسعه یافته است بنابراین ضرورت دارد تا به طور مختصر به این علم و نقش و اهمیت آن بپردازیم.

آواشناسی همواره نقش بسزایی در علم زبان‌شناسی داشته است به طوری که بسیاری از زبان‌شناسان پیش از مکتب گشتاری، آواشناس بوده اند. « زبان‌شناسان قبلی همه در وهله اول آواشناس بودند. تقریباً همه زبان‌شناسان اروپایی دردرجه اول آواشناس بوده و هستند.» [۵].

اهمیت آواشناسی از آنجا مشخص می‌گردد که این علم به ما کمک می‌کند تا با همه جنبه‌های صوتی زبان آشنا گردیم. این علم چگونگی استفاده انسان از اندام‌های صوتی جهت تولید آواهای زبانی، چگونگی دریافت آواها و همچنین خصوصیات و مختصات فیزیکی آواها را مورد مطالعه قرار می‌دهد. یکی از موارد استفاده علم آواشناسی این است که دانش‌آموزان و مدرسان زبان می‌توانند با توجه مباحث مطرح در این علم در خصوص چگونگی تولید آواها در هر زبان جهت بهتر تلفظ کردن لغات استفاده نمایند. کیسلینگ[1] [۶] معتقد به این امر می‌باشد که جلب توجه زبان‌آموزان به ویژگی آوایی آواهای زبان مورد نظر، هر چند به طور خلاصه، مناسب‌تر از این است که آنها را صرفاً در معرض زبان دوم قرار دهیم به امید اینکه آنها خود قادر به کشف ویژگی‌های آوایی آواهای مورد نظر را باشند.

آواشناسی همچنین با مشخص کردن مشخصه‌های هر آوای گفتاری به جامعه‌شناسان زبان این امکان را می‌دهد تا بتوانند به طور دقیق‌تر و موثرتری به مطالعه یک زبان بخصوص بپردازند و به این طریق درک بهتری از سبک‌ها و لهجه‌های آن زبان داشته باشند. نقش آواشناسی در توانبخشی گفتاری و شنواسنجی بسیار مهم می‌باشد زیرا که آواشناسی در پی‌یافتن چگونگی تولید و دریافت آواهای زبانی و همچنین مشخص کردن ویژگی‌های هر کدام از آواهای گفتاری و بکارگیری صحیح آنها می‌باشد. آواشناسی با تعریف مشخصه‌های فیزیکی

1 Elizabeth M. Kissling

آواهایی هر زبان این امکان را بوجود می‌آورد که بتوان آواهای الکترونیکی متناسب با آواهای آن زبان را بوجود آورد که این آواهای الکترونیکی در زمینه‌های مختلفی اعم از تجارت و صنعت، آموزش و پرورش و صنعت سرگرمی قابل استفاده می‌باشد. پزشک قانونی نیز می‌تواند با استفاده از طیف نگارهای صوتی که با استفاده از علم آواشناسی ساخته شده‌اند، هر چه بهتر مجرمان را شناسایی نماید. آواشناسی همچنین می‌تواند در علوم دیگر نظیر روانشناسی، علوم رفتاری، علوم شناختی، فناوری‌های زبانی، علوم کامپیوتری، فیزیک، پزشکی، صنعت مخابرات مورد استفاده قرار گیرد.

با توجه به آنچه در خصوص اهمیت علم آواشناسی گفته شد ما به ابزاری نوین نیاز داریم که در فهم بیشتر این علم ما را یاری نماید. از آنجا که علم هستان‌شناسی ابزاری کارآمد می‌باشد که بوسیله آن می‌توان از یک دامنه خاص واقعی یا فرضی مدلی از دانش ایجاد کرد و به عنوان یک سیستم نمایش اطلاعات روابط بین مفاهیم و اشیاء را مشخص نمود و همچنین در بازیابی، ذخیره و اشتراک گذاری اطلاعات ابزار بسیار مناسبی می‌باشد، ما بر آن شدیم که با توجه به عدم وجود یک هستان‌شناسی مناسب در زبان فارسی جهت شناساندن علم آواشناسی و مشخص کردن روابط بین مفاهیم و واژگان و همچنین مصور سازی مفاهیم علم آواشناسی به ایجاد هستان‌شناسی آواشناسی بپردازیم.

هستان‌شناسی آواشناسی به مباحثی از قبیل مفاهیم مرتبط به آواهای زبانی انسان، دسته‌بندی آواها، مفاهیم به کار گرفته شده در شاخه‌های سه‌گانه علم آواشناسی، ابزارهای به کار گرفته شده در این علم و اندام‌های درگیر در تولید آواهای زبانی می‌پردازد.

با توجه به حوزه موضوعی انتخاب شده، این پژوهش اهداف زیر را دنبال می‌نماید:

۱) شناسایی مفاهیم و واژگان آواشناسی و استخراج آنها
۲) یافتن روابط موجود بین مفاهیم و اصطلاحات آواشناسی
۳) یافتن نرم‌افزار مناسب جهت پیاده‌سازی هستان‌شناسی آواشناسی
۴) پیاده‌سازی و مصورسازی نرم افزاری مفاهیم و واژگان آواشناسی و رابطه میان آنها
۵) توسعه مفاهیم و واژگان آواشناسی در شبکه جهانی وب در قالب اچ.تی.ام.ال

۲- روش ساخت هستان‌شناسی آواشناسی در زبان فارسی

ابتدا این نکته را باید خاطر نشان کرد که پژوهش حاضر با استفاده از ابزار ساخت

هستان‌شناسی پروتژه [۷] نسخه۰,۲,۵، پیاده‌سازی شده است. این ابزار ساخت هستان‌شناسی به صورت واسط کاربر عمل کرده و محیطی با نمایه‌های متعدد خود فراهم میآورد که به توسعه‌دهندگان هستان‌شناسی این امکان را می‌دهد تا با توجه به نیازهای مورد نظر خود از هر یک از نمایه‌های موجود جهت ساخت هستان‌شناسی‌های مطلوب خود بهره گیرند. این هستان‌شناسی نیز با توجه به هدف تعیین شده خود از نمایه‌های موجود در ابزار ساخت هستان‌شناسی پروتژه که مورد نیاز بوده‌اند استفاده نموده است.

پروژه توسعه هستان‌شناسی آواشناسی شامل مراحل تعیین کلاس‌ها و سلسله مراتب آنها، تعیین ویژگی کلاس‌ها، تعریف نمونه‌ها، حاشیه‌نویسی برای کلاس‌ها و استنتاج هستان‌شناسی است و در مرحله آخر و مجزا از مراحل قبل هستان‌شناسی آواشناسی به زبان اچ.تی.ام.ال تبدیل گردید تا به این طریق علاقه‌مندان و افراد فعال در زمینه آواشناسی به این هستان‌شناسی دسترسی آسان‌تری داشته باشند. در ادامه هر یک از مراحل ساخت هستان‌شناسی آواشناسی توضیح داده خواهد شد.

۲-۱ تعیین کلاس‌ها و سلسله مراتب آنها

در انجام این پژوهش ابتدا به تعداد ۳۷۷ مفهوم و اصطلاح آواشناسی با استفاده از منابع فارسی و انگلیسی شناسایی و استخراج.گردید که این منابع شامل تألیفات حق‌شناس [۸]، ثمره[۹] ، بی‌جن‌خان [۱۰]، مشکوةالدینی [۱۱] و کریستال [۱۲] است. سپس واژه‌ها و اصطلاحات آواشناسی استخراج شده به عنوان کلاس‌های هستان‌شناسی تعیین گردیدند و ساختاری سلسله مراتبی از آنها به وسیله نمایه سلسله مراتب کلاس‌های پروتژه ایجاد گردید. نمایه سلسله مراتب کلاس‌های شامل دو دسته کلاس مجزای، کلاس‌های اعلانی و کلاس‌های استنتاجی می‌باشد که هر یک توسط زبانه‌های[1] خاص خود امکان نمایش پیدا می‌کنند. از میان دو نوع کلاس موجود، کلاس اعلانی برای ایجاد ساختار سلسله مراتبی از مفاهیم و اصطلاحات آواشناسی انتخاب گردید. این نمایه به ما کمک نمود تا یک ساختار سلسله مراتبی از مفاهیم و اصطلاحات آواشناسی را با استفاده از گزینه‌ای به نام SubClassOF ایجاد نماییم. سلسله مراتب کلاس‌های ایجاد شده یک ساختار درختی را موجب گردید که در آن هر کلاس دارای

Tabs1

زیر کلاس‌هایی مرتبط به خود است. به عبارتی دیگر گره‌های فرزند موجود در درخت در زیرگره‌های والد مربوط به خود جای گرفته است.

در هستان‌شناسی آواشناسی، زبان‌شناسی اولین کلاس ایجاد شده می‌باشد که کلاس آواشناسی به عنوان زیرکلاس آن واقع گردیده است. کلاس آواشناسی به نوبه خود دارای چندین زیرکلاس از جمله آواشناسی فیزیکی، آواشناسی شنیداری، آواشناسی تولیدی و ..... می‌باشد و هر یک از این زیرکلاس‌های آواشناسی خود تبدیل به کلاس‌هایی گردیده‌اند که دارای زیرکلاس‌های مختص خود می‌باشند و به این طریق طبق یک فرایند تکراری حول محور کلاس آواشناسی این هستان‌شناسی ایجاد گردید.

۲-۲ تعیین ویژگی کلاس‌ها

تعیین ویژگی‌ها در هستان‌شناسی آواشناسی به این صورت انجام گرفت که بعد از تهیه فهرستی از مفاهیم و اصطلاحات آواشناسی، اقدام به یافتن روابط بین مفاهیم و اصطلاحات با استفاده از کتاب «فرهنگ لغت زبانشناسی و آواشناسی»، «تزاروس دات کام » [۱۳] و هستان‌شناسی «گلد[1]» [۱۴] گردید و مبنا بر این قرار گرفت که از چهار ویژگی «نوعی از[2]»، «جزئی از[3]»، «عضوی از[4]» و «مرتبط با[5]» برای بیان روابط بین مفاهیم و اصطلاحات آواشناسی استفاده گردید. سپس پرسش‌نامه‌ای از واژه‌های آواشناسی و ارتباط آنها با یکدیگر در قالب چهار ویژگی «نوعی از»، «جزئی از»، «عضوی از» و «مرتبط با» تهیه گردید و در اختیار متخصصین و اساتید این حوزه قرار گرفت که تا میزان زیادی روابط ایجاد شده را تایید نمودند و در مواردی که نیاز به اصلاح بود نظرات ایشان اعمال گردید. در مرحله آخر ویژگی‌ها در نمایه ویژگی شئ پروتژه وارد گردید و سپس با استفاده از نمایه‌های دامنه و محدوده اقدام به تعریف رابطه بین کلاس‌های آواشناسی گردید. در زیر چهار ویژگی‌های به کار گرفته شده در هستان‌شناسی آواشناسی تعریف گردیده است.

ویژگی جزئی از: یکی از ویژگی‌هایی است که در هستان‌شناسی آواشناسی بکار گرفته

1 General Ontology for Linguistic Description
2 Type of
3 Part of
4 Member of
5 Related to

شده است. این ویژگی روابط مفهومی ایجاد می‌کند که بوسیله آن می‌توان نوعی از رابطه سلسله مراتبی ایجاد نمود و رابطه‌ی کل به جزء را میان مفاهیم و اصطلاحات آواشناسی را نشان داد. به عنوان مثال در هستان‌شناسی آواشناسی مشخصه «زبرزنجیری» به عنوان جزئی از «مشخصه آوایی» نام برده شده است.

ویژگی عضوی از: این ویژگی روابط مفهومی را موجب می‌گردد که به واسطه آن می‌توان رابطه‌ی میان یک واژه و یا یک اصطلاح را به مثابه یک عضو از یک مجموعه را نشان داد. در این باره می‌توان مثلاً از رابطه‌ی «حنجره» نسبت به «اندام تولید» در هستان‌شناسی آواشناسی نام برد. در مثال آورده شده حنجره به عنوان عضوی از مجموعه اندام‌های تولید می‌باشد.

ویژگی مرتبط با: یکی از روابط به کار گرفته شده در هستان‌شناسی آواشناسی می‌باشد که واژگان و اصطلاحاتی را که دارای ویژگی‌های مشترک با یکدیگر می‌باشند را به هم متصل می‌کند. به عنوان مثال در هستان‌شناسی آواشناسی «آواسازی» واژه‌ای است که با «تارآوا مرتبط» خوانده شده است زیرا آواسازی به لرزش و ارتعاش تارآواها اشاره دارد.

ویژگی نوعی از: این ویژگی مشخص کننده رابطه‌ای است که مفهوم یک واژه یا اصطلاح مفهوم واژه یا اصطلاح دیگری را در برمی‌گیرد، به عنوان مثال در هستان‌شناسی آواشناسی لبی‌شدگی نوعی از تولید دومین محسوب شده است. در رابطه ایجاد گردیده به وسیله این ویژگی، تولید دومین به عنوان کلاسی می‌باشد که لبی‌شدگی را به عنوان زیرکلاس در خود جای داده است. البته باید به این نکته توجه داشت که واژه لبی‌شدگی از شرایط لازم و کافی بیشتری نسبت به واژه تولید دومین برخوردار می‌باشد.

۲-۳ تعریف نمونه‌ها

نمونه‌ها موارد واقعی هستند که در یک هستان‌شناسی به کار گرفته می‌شوند. نمونه‌های به کار گرفته شده در هستان‌شناسی آواشناسی از کتاب‌های حق‌شناس، ثمره، و مشکوةالدینی استخراج گردیده است. برای پیاده‌سازی نمونه‌ها در ابزار ساخت هستان‌شناسی پروتژه ابتدا کلاس مناسب انتخاب گردید و سپس آن نمونه برای آن را کلاس وارد گردید. در هستان‌شناسی آواشناسی ۳۹ نمونه به کار گرفته شده است. در جدول شماره ۱ نمونه‌ای از نمونه‌های به کارگرفته شده در هستان‌شناسی آواشناسی آورده شده است.

جدول شماره ۱ نمونه‌های به کار رفته در هستان‌شناسی آواشناسی

نمونه‌های به کار گرفته شده در هستان‌شناسی آواشناسی								
a	b	d	e	ei	f	h	i	j
l	m	n	o	ou	p	q	r	s
u	v	w	x	y	z	Ʒ	K	t

۳-۲ حاشیه‌نویسی برای کلاس‌ها

حاشیه نویسی یکی دیگر از مراحل توسعه هستان‌شناسی می‌باشد. در مرحله حاشیه‌نویسی اطلاعاتی در هستان‌شناسی مورد توسعه وارد می‌گردد اما این اطلاعات هیچ گونه تاثیری در استنتاج هستان‌شناسی مورد نظر ندارد بلکه به این طریق توسعه‌دهنده هستان‌شناسی قادر خواهد بود تا مراحل توسعه و اجزای درونی هستان‌شناسی را توضیح دهد. در هستان‌شناسی آواشناسی به تعداد ۳۷۰ حاشیه‌نویسی صورت گرفته است که در خلال آن به تعریف واژه‌ها و اصطلاحات آواشناسی پرداخته گردیده است.

۳-۲ استنتاج هستان‌شناسی

مرحله استنتاج هستان‌شناسی نقش بسیار مهمی در ایجاد یک هستان‌شناسی صحیح و سازگار بازی می‌کند. در این مرحله سازگاری هستان‌شناسی بررسی گردیده و سپس تناقضات منطقی را که در تعاریف ذکر شده‌اند را آشکار می‌گردد. آزمون سازگاری دانش شامل شناسایی بازخورد، انتقال و حشو دانش است. در نرم افزار پروتژه از ابزار هرمیت[1] استفاده می‌شود تا سازگاری هستان‌شناسی بررسی گردد و همچنین روابط وابستگی بین کلاس‌ها مشخص گردد. در هستان‌شناسی آواشناسی نیز از این ابزار استنتاجی استفاده شد و هیچگونه ناسازگاری مشاهده نگردید.

۳- ابزار ساخت هستان‌شناسی پروتژه

همان طور که پیش از این هم گفته شد هستانشناسی آواشناسی با استفاده از ابزار ساخت هستان‌شناسی پروتژه ایجاد گردیده است. این ابزار هستان‌شناسی به این دلیل انتخاب

1 HermiT

گردیده‌است زیرا که انعطاف‌پذیری خوبی را برای مدل‌سازی بهینه فراهم می‌کند. پروتژه همچنین این امکان را بوجود می‌آورد تا بتوان هستان‌شناسی‌هایی را با دامنه موضوعی خاص پدید آورد. این ابزار ساخت هستان‌شناسی دارای مدل دانش انعطاف‌پذیرو معماری افزون‌هایی قابل گسترش است. پروتژه عمدتا از ایجاد و مصورسازی حاشیه نویسی پشتیبانی می‌کند. این ابزار ساخت هستان‌شناسی نه تنها اجازه می‌دهد تا توسعه دهندگان به ارائه نمایش مدل داخلی بپردازند، بلکه به طور آزادانه رابط کاربری را با توجه به نیاز شخصی خود ایجاد نمایند. یکی از نقاط قوت آن این است که به طور خودکار می‌تواند یک رابط کاربری را از تعاریف کلاس‌ها تولید کند و بنابراین اکتساب دانش را با سرعت خوبی پشتیبانی کند[۱۵].

۴- پیاده‌سازی هستان‌شناسی آواشناسی در قالب اچ.تی.ام.ال

پیاده‌سازی هستان‌شناسی آواشناسی به زبان اچ.تی.ام.ال به این منظور صورت گرفت تا اینکه هستان‌شناسی آواشناسی این قابلیت را داشته باشد تا آن را در محیط وب قرار داد و به این طریق علاقه‌مندان و افراد بیشتری به داده‌های موجود در این هستان‌شناسی دسترسی داشته باشند. پیاده‌سازی هستان‌شناسی آواشناسی در قالب اچ.تی.ام.ال با تگ‌های مختص این زبان صورت گرفت. شکل ۱ در پیوست صفحه وب ایجاد شده هستان‌شناسی آواشناسی را نشان می‌دهد که در آن فهرست واژه‌های آواشناسی آورده شده است. در این صفحه با کلیک بر روی واژه مورد نظر می‌توان همان‌طور که در شکل ۲ پیوست آورده شده است تعریف و نوع ارتباط آن با واژه‌های دیگر آواشناسی را مشاهده نمود.

۵- نتیجه‌گیری

در این هستان‌شناسی آواشناسی ۳۷۷ اصطلاح با استفاده از نرم‌افزار ساخت هستان‌شناسی پیاده‌سازی گردید و سپس هستان‌شناسی ایجاد گردیده با دو هستان‌شناسی فارس‌نت و گلد مقایسه گردید.

فارس نت به این دلیل انتخاب گردید زیرا که در زبان فارسی به حوزه هستان‌شناسی‌های موضوعی، خصوصاً زبان‌شناسی توجه چندانی نشده است تا مبنای مقایسه با هستان‌شناسی آواشناسی قرار گیرد بنابراین ما به هستان‌شناسی عمومی فارس‌نت مراجعه نموده و این

هستان‌شناسی را در زمینه مفاهیم و واژگان آواشناسی بررسی نمودیم. بررسی مورد نظر به این گونه انجام شد که تعدادی از واژگان آواشناسی به صورت تصادفی انتخاب گردیدند و در این هستان‌شناسی مورد جستجو قرار گرفتند . با توجه به موارد مورد جستجو این نتیجه بدست آمد که فارس‌نت تنها ۱۴/۵ از مفاهیم و واژگان آواشناسی شامل می‌گردد و مفاهیم و واژگان تخصصی آواشناسی را در برنمی‌گیرد و همچنین این نکته آشکار گردید که هستان‌شناسی فارس‌نت تنها به تعیین روابط در سطح عمومی واژگان و مفاهیم پرداخته می‌شود. به عنوان مثال در تعیین ویژگی واژه آواشناسی «رهش» از دو ویژگی hypernym و related to استفاده گردیده است. در قسمت بیان ویژگی hypernym برای رهش همان طور که در شکل ۱ مشاهده می‌شود تنها به ویژگی‌های عمومی اکتفا کرده و ارتباط آن با آواشناسی مشخص نگردیده است و در قسمت ویژگی related to واژه رهش مرتبط با همخوان، خیشومی و انسدادی آورده شده است که بیان دقیقی برای توصیف این واژه در آواشناسی نمی‌باشد. جدول ۲ نمونه‌ای از واژه‌های مورد بررسی آورده شده است.

شکل ۱: نمایش ویژگی hypernym برای واژه آواشناسی رهش در هستان‌شناسی فارس‌نت

جدول ۲: تعداد واژه‌های مورد بررسی در هستان‌شناسی فارس‌نت

زنشی	جایگاه تولید	شیوه تولید
فورانی	واک‌رفته	ناسوده
لبی شدگی	واک نفسی	انسدادخیشومی
تودماغی	نشانه مضاعف	کامی شدگی
واکه مرکب	نادمیده	تیغه زبان
ساخت‌نوفه‌ای	هم‌تولیدی	واک نفسی
آغازه غلت	غلطان	رسا
افتان	جایگاه واک‌ساز	مشخصه آوایی

هستان‌شناسی مطرح گلد که متعلق به زبان‌شناسی توصیفی است یکی دیگر از هستان‌شناسی‌ها می‌باشد که مورد مقایسه با هستان‌شناسی آواشناسی قرار گرفت. هستان‌شناسی گلد شامل ۵۰۳ واژه زبان‌شناسی می‌باشد. آنچه قابل ملاحظه می‌باشد هستان‌شناسی گلد از هیچ نمونه‌ای برای کلاس‌های خود استفاده ننموده است در حالی که هستان‌شناسی آواشناسی در تعریف کلاس‌های خود از ۳۹ نمونه استفاده نموده است. تفاوت دیگر هستان‌شناسی آواشناسی با هستان‌شناسی گلد این می‌باشد که در هستان‌شناسی آواشناسی از چهار ویژگی برای بیان روابط کلاس‌ها استفاده گردیده است در حالی که هستان‌شناسی گلد ۷۶ ویژگی برای تعریف کلاس‌های خود استفاده نموده است که البته این امر باعث گردیده است تا نتواند به طور دقیق ارتباط بین اجزاء هستان‌شناسی را به عنوان یک مجموعه منسجم ارائه نماید.

منابع

[1] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing?. *International journal of human-computer studies*, *43*(5-6), 907-928.

[2] Gómez-Pérez, A., & Benjamins, R. (1999). Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods. IJCAI and the Scandinavian AI Societies. CEUR Workshop

Proceedings.

[3] Uschold, M., & King, M. (1995). *Towards a methodology for building ontologies* (pp. 15-30). Edinburgh: Artificial Intelligence Applications Institute, University of Edinburgh.

[4] Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them?. *IEEE Intelligent Systems and their applications*, *14*(1), 20-26.

[۵] باطنی، محمدرضا (۱۳۸۰). *زبان و تفکر*. تهران: آبانگاه.

[6] Kissling, E. M. (2013). Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners?. *The modern language journal*, *97*(3), 720-744.

[7] https://protegewiki.stanford.edu/wiki/Main_Page

[۸] حق شناس، علی‌محمد (۱۳۹۲). *آواشناسی*. تهران: آگه.

[۹] ثمره، یدالله (۱۳۷۸). *آواشناسی زبان فارسی آواها و ساخت آوایی هجا*، تهران: مرکز نشر دانشگاهی.

[۱۰] بی‌جن‌خان، محمود (۱۳۹۲). *نظام آوایی زبان فارسی*، تهران: سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌های (سمت).

[۱۱] مشکوةالدینی، مهدی (۱۳۹۳). *سخت آوایی زبان*، مشهد: دانشگاه فردوسی مشهد.

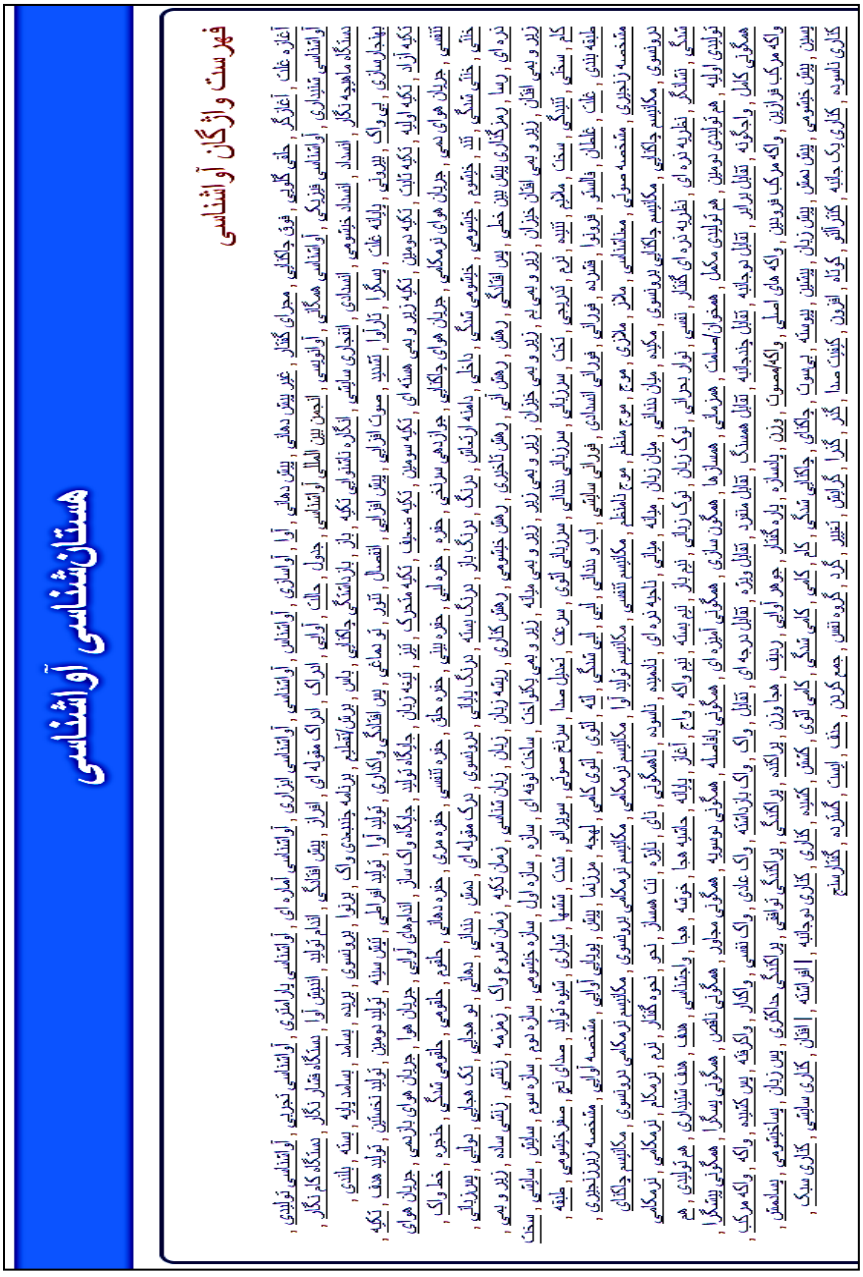
[12] Crystal, D. (2012). *A dictionary linguistic and phonetics*. New Jersey: Blackwell.

[13] http://www.thesaurus.com/

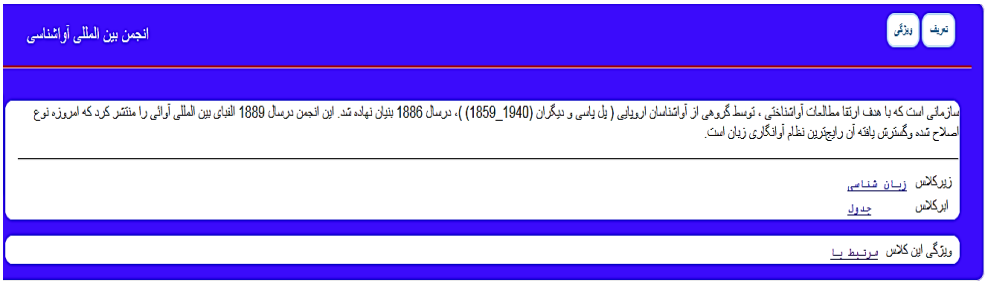
[14] *About Gold*. Retrieved May 1, 2017, from http://linguistics-ontology.org//info/about.

[15] Knublauch, H., & Musen, M. A. (2004, June). Weaving the biomedical semantic web with the Protégé OWLplugin. In *Proceedings of the First International Conference on Formal Biomedical Knowledge Representation-Volume 102* (pp. 39-47). American Medical Informatics Association.

پیوست



شکل ۱پیوست صفحه وب ایجاد شده واژه‌ها و اصطلاحات آواشناسی در هستان‌شناسی آواشناسی



شکل ۲ پیوست چگونگی نمایش واژه‌های هستان‌شناسی آواشناسی با تعریف و ویژگی آنها

هستی‌شناسی و بازیابی اطلاعات مطالعه موردی: حوزه ریاضیات

شبنم رشیدی تبار* ، **فرامرز سهیلی**** و **مریم فیضی*****

چکیده

هستی‌شناسی ابزاری برای نمایش رسمی و به اشتراک گذاشتن دانش حوزه‌ای خاص از طریق مدل‌سازی و ایجاد چارچوبی از مفاهیم و روابط معنایی بین آنهاست. هستی-شناسی در حوزه ریاضیات مانند سایر حوزه‌ها دارای کاربرد است. در این پژوهش، به منظور پیشنهاد روشی برای ساخت هستی‌شناسی حوزه ریاضیات و بسط و گسترش بسیار وسیع‌تر آن در آینده به زبان فارسی، به ایجاد نمونه اولیه آن اقدام شد. روش پژوهش حاضر از نوع توصیفی- توسعه‌ای است و برای ایجاد هستی‌شناسی ریاضیات با استفاده از اصطلاحنامه ریاضیات دامنه موضوع به‌دست آمد و مفاهیم و روابط مربوط به آن‌ها از متون و منابع استخراج و برای طراحی هستی‌شناسی نیز از نرم‌افزار پروتژ نسخه ۵ (۱، ۵، ۳) استفاده شد. پس از شناسایی حوزه موضوعی ریاضی، چهار مرحله به عنوان مراحل اصلی ساخت هستی‌شناسی ریاضی شناخته شد. این مراحل شامل: شناسایی دامنه و حوزه موضوعی آنتولوژی، تعریف کلاس‌ها و ساختار آن‌ها، تعریف ویژگی‌های کلاس‌ها و روابط و ایجاد نمونه‌ها بود.با توجه به یافته‌های این پژوهش، روش ارائه شده در این پژوهش می‌تواند در گسترش هستی‌شناسی حوزه ریاضیات در زبان فارسی مفید واقع شود.

واژه‌های کلیدی: هستی‌شناسی، روش‌شناسی ایجاد هستی‌شناسی، ریاضیات.

۱. مقدمه

دسترس پذیری، سهولت یافتن و رؤیت اطلاعات مرتبط ارزش افزوده‌ای است که از طریق نظام‌ها یا اطلاعاتی مطلوب حاصل می‌آید. هدف شبکه جهانی وب نیز مانند هر نظام اطلاعاتی

* کارشناسی ارشد علم اطلاعات و دانش شناسی و کارشناس کتابخانه‌های عمومی استان کردستان(نویسنده مسئول) shabnamrashidi61@yahoo.com

** دانشیار گروه علم اطلاعات و دانش شناسی دانشگاه پیام نور کرمانشاه fsohieli@gmail.com

*** کارشناس ارشد علم اطلاعات و دانش‌شناسی m.feyzi313@gmail.com

دسترسی سریع به منابع مرتبط است. با ایجاد و فراگیر شدن وب، بازیابی و رتبه‌بندی منابع توسط موتورهای جستجو از مهم‌ترین مسائل مورد توجه کاربران و ارائه‌دهندگان خدمات این شبکه جهانی بوده است [۱]. محیط وب با ساختار گسترده، امکان دسترسی به اطلاعات وسیعی را برای کاربران خود فراهم آورده است. اهمیت توجه به امر سازماندهی و بازیابی اطلاعات در محیط وب و تغییر انتظارهای کاربران از این شبکه جهان‌گستر، تلاش برای ارتقای آن را ضروری ساخته است، زیرا از یک سو، سهولت دسترسی به منابع وبی، همگان را به استفاده از آن فرا می‌خواند و از سوی دیگر حجم وسیع، متنوع و غیرقابل مدیریت آن، مسائلی را در خصوص بازیابی اطلاعات در این محیط مطرح می‌سازد [۲]. وب معنایی که حاصل تلاش کنسرسیوم جهانی وب است، به منظور اشتراک اطلاعات در وب و جستجو بر اساس موضوع و ارتباط میان داده‌ها ایجاد شده است و نه تنها برای انسان قابل فهم است، بلکه ماشین‌ها نیز توانایی فهم آن را دارند [۳]. هدف وب معنایی که به عنوان وب داده‌ها نیز شناخته می‌شود، یکپارچه‌سازی داده‌های منابع گوناگون است [۴]. اما برای توسعه مطلوب وب معنایی و افزایش کیفیت سازماندهی و در نتیجه بازیابی اطلاعات در این محیط، طراحی ابزارهای مناسب از جمله هستی‌شناسی‌ها ضروری است [۵].

هستی‌شناسی[1]ها برای علاقه‌مندان به تکامل مستمر وب، و به‌ویژه افراد فعال در توسعه وب معنایی، موضوع هیجان‌انگیز و پرطرفداری است. هر چند هستی‌شناسی‌ها در حال حاضر موضوعی عمومی شده‌اند، اما با این وجود هنوز ابهامات زیادی درباره آنها وجود دارد [۶]، هستی‌شناسی‌ها در محدوده وسیعی کاربرد دارند. از جمله می‌توان به شبکه‌های جهان‌گستر معنایی، موتورهای جستجو، تجارت الکترونیکی، پردازش زبان طبیعی، مهندسی دانش، استخراج و بازیابی اطلاعات، کتابخانه‌های رقمی و مانند آنها اشاره کرد [۷]. محبوبیت هستی شناسی‌ها (هستی‌شناسی در وب معنایی اصطلاحات و ارتباط بین آن‌ها در دامنه‌ی مفروض را بیان می‌کند [۸]، از زمان ظهور وب معنایی به سرعت در حال گسترش است، زیرا برای استفاده از وب معنایی، اطلاعات باید به فرمت قابل فهم برای ماشین نمایش داده شوند. کنسرسیوم جهانی وب W3C چند فرمت برای نمایش داده‌ها در وب معنایی از جمله،RDF، Schema RDF، OWL پیشنهاد کرده است [۹]. RDF نخستین زبانی است که توسط کنسرسیوم

1 .Ontology

جهانی وب، منحصرا برای وب معنایی ارائه شده است و یک زبان همه منظوره برای توصیف منابع بوده و به‌طور اخص برای بیان فراداده‌هایی در مورد داده‌های موجود در وب، استفاده می‌شود. به‌طور کلی این زبان این قابلیت را داراست که در مورد هر موجود یا شی شناسه دار، اطلاعاتی در قالب گزاره‌ها بیان کند. زبان RDF مستقل از دامنه بوده و قابلیت توصیف آن را ندارد. با RDF می‌توان در مورد یک موجود شناسه دار گزاره‌هایی بیان نمود ولی نمی‌توان گفت آن موجود چیست. دلیل این ویژگی این است که RDF قدرت توصیف هستی‌شناسی را ندارد و با استفاده از هستی‌شناسی است که می‌توان یک دامنه و مفاهیم و موجودات آن را به طور مناسب توصیف کرد. برای مرتفع کردن این کمبود، RDF Schema به عنوان یک راه حل مطرح است و زبان تعریف RDF است و آن را به گونه‌ای توسعه می‌دهد که برخی از نیازهای کاربران برای تعریف کلاس‌ها و خصیصه‌ها را در یک نظام سلسله مراتبی رفع می‌کند. RDF Schema در حقیقت واژگان RDF را توسعه می‌دهد. ولی در اغلب موارد نیازمندی‌های موجود در فرآیند ایجاد یک هستی‌شناسی، فراتر از توانمندی‌های RDF-S است. به عنوان مثال RDF-S این قابلیت را ندارد که بیان کند یک کلاس خاص با کلاس دیگر معادل است. بنابراین می‌توان ادعا کرد RDF-S یک زبان ضعیف برای توصیف هستی‌شناسی است. OWL در سال ۲۰۰۴ توسط کنسرسیوم جهانی وب ارائه گردید. این زبان یک زبان مبتنی بر RSD(S)[1] است و تمام کاستی‌های زبان‌های قبلی را رفع کرده و به عنوان زبان تمام عیار برای توصیف هستی‌شناسی‌ها مطرح است. و مانند RSD(S) دارای یک واژگان است ولی از قابلیت‌ها بسیار غنی‌تری برخوردار است [۱۰]. بررسی متون حاکی از آن است که پژوهش‌های فراوانی راجب به هستی‌شناسی صورت گرفته، اما پژوهش در داخل کشور درباره ساخت آنتولوژی رشته‌ها و حوزه‌های مختلف محدود بوده است. در این مقاله، روش به کار رفته برای ساخت هستی‌شناسی ریاضی شرح داده می‌شود. هدف پژوهش ارائه روش‌شناسی برای طراحی هستی‌شناسی‌های حوزه ریاضی در زبان فارسی است.

مرور پژوهش‌ها نشان‌دهنده این است که پژوهش‌های زیادی راجب به هستی‌شناسی‌ها صورت گرفته است. پژوهش‌ها درباره ساخت هستی‌شناسی‌ها، ابزارهای ساخت هستی‌شناسی، زبان‌های

1. RDF and RDF-S

هستی‌شناسی و موارد بسیار دیگر گویای این واقعیت است. خارج از کشور، پژوهش ژو[1] و همکاران [۱۱] درباره ایجاد هستی‌شناسی، پژوهش رویزمارتینز[2] و همکاران [۱۲] درباره ایجاد هستی شناسی برای زیست شناسی، پژوهش‌های برایت[3] و همکاران [۱۳] درباره ایجاد هستی-شناسی برای راهنمایی در تجویز آنتی‌بیوتیک و پژوهش سیامایلی و ریکا درباره توسعه هستی-شناسی اساطیر یونانی [۱۴] و در داخل کشور نیز پژوهش شمس‌فرد [۱۵]، فتحیان [۱۶]، زاهدی و همکاران [۱۷].

۲. روش پژوهش

روش پژوهش حاضر از نوع توصیفی- توسعه‌ای است و برای ایجاد هستی‌شناسی ریاضیات با استفاده از اصطلاحنامه ریاضیات [۱۸] دامنه موضوع به‌دست آمد و مفاهیم و روابط مربوط به آن‌ها از متون و منابع استخراج و برای طراحی آنتولوژی نیز از نرم‌افزار پروتژه[4] نسخه ۵ (۱، ۵، ۳) استفاده شد.

پروتژ، نرم افزاری برای ایجاد و ویرایش هستی‌شناسی‌ها و پایگاه‌های دانش است این نرم-افزار رابط کاربری فراهم می‌آورد که در آن امکان تعریف مفاهیم، نمونه‌ها، ویژگی‌ها و محدودیت‌های مفاهیم و هم‌چنین روابط، وجود دارد [۱۹]. این نرم‌افزار یک نرم‌افزار متن باز است که به صورت قیاسی طبقه‌بندی گرافیکی خود را ارائه می‌دهد و مبتنی بر زبان جاوا است که در سال ۱۹۸۷ توسط دانشگاه استنفورد آمریکا ساخته شد. هدف اصلی آن هم ساده‌سازی فرایند اکتساب دانش برای سیستم‌های خبره بیان شده است [۲۰].

۳. مراحل طراحی هستی‌شناسی ریاضی به وسیله نرم‌افزار پروتژ

در این قسمت به شرح مراحل پیموده شده می‌پردازیم:

۱-۴ . شناسایی دامنه و حوزه موضوعی هستی‌شناسی:

حوزه موضوعی این پژوهش ریاضیات است و مبنای آن مفاهیم و روابط معنایی است که از

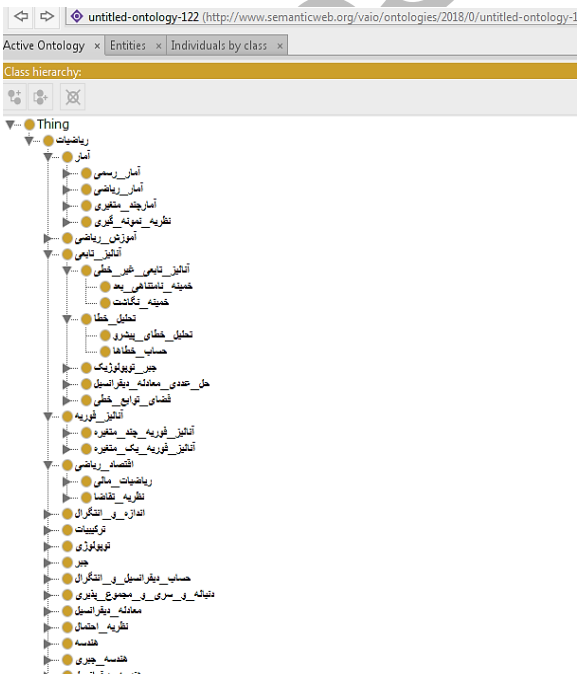
---

1. Zhou
2. Ruiz-Martínez.
3. Bright
4. Protégé

متون مرتبط با ریاضی استخراج شده‌اند. این متون شامل کتاب‌های ریاضیات چیست؟ [۲۱]، تئوری مقدماتی اعداد [۲۲]، هندسه، زوایا، اعداد [۲۳]، نظریه گراف و کاربردهای آن [۲۴]، آنالیز فوریه کاربردی [۲۵]، ریاضیات عمومی و کاربردهای آن [۲۶]، اقتصاد ریاضی [۲۷] می‌باشد.

۲-۴ . تعریف کلاس‌ها و ساختار آن‌ها:

در این مرحله کلاس‌هایی که از طریق اصطلاحنامه به‌دست آمده است را به طریق سلسله - مراتبی از کل به جزء وارد نرم افزار می‌کنیم. در این پژوهش موضوع ریاضی با کلاس‌هایی از جمله آمار، آنالیز تابعی، آموزش ریاضی، آنالیز فوریه، آنالیز همساز مجرد، اقتصاد ریاضی، ترکیبات، توپولوژی، حساب وردش، دنباله و سری مجموع‌پذیری، جبر و حساب دیفرانسیل و انتگرال و غیره و زیرکلاس‌های هر کلاس نیز وارد نرم‌افزار گردید. در شکل ۱ نمونه‌ای از کلاس‌ها و زیر کلاس‌های وارد شده در نرم‌افزار مشاهده می‌گردد.



شکل ۱. بخش تعریف کلاس‌ها و زیر کلاس‌ها

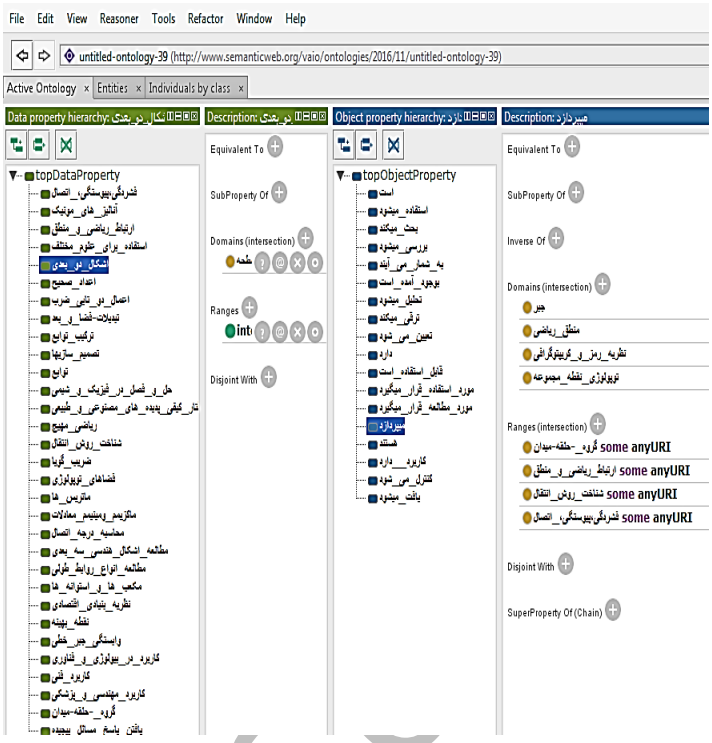
۳-۴. تعریف ویژگی‌های[1] کلاس‌ها و روابط:

اطلاعات ارائه شده در کلاس‌ها کافی نیست به همین منظور، قسمت ویژگی‌ها به روابط مورد نیاز در هستی‌شناسی اختصاص می‌یابد و ارتباط رده‌ها از این طریق ایجاد می‌شود. دو نوع اصلی از ویژگی وجود دارد، ویژگی‌های نوع شئ[2] و ویژگی‌های نوع داده[3].

ویژگی‌های نوع شئ، ارتباط بین دو کلاس، دو نمونه یا دو شئ را بیان می‌کند و ویژگی-های نوع داده ارتباط بین یک نمونه یا شئ با مقادیر داده را مشخص می‌کند. هر ویژگی یک دامنه۴ و یک برد۵ دارد که می‌توان گفت یک ویژگی عناصر موجود در دامنه خود را به عناصر موجود در برد خود مرتبط می‌کند، دامنه عبارت است از مجموعه عناصری که آن ویژگی به آن تعلق می‌گیرد و برد عبارت است از مجموعه عناصری که به عنوان مقادیر آن ویژگی می‌تواند استفاده شود. به عنوان مثال در هستی شناسی ریاضیات، یک ویژگی نوع داده به نام "اشکال دو بعدی" می‌سازیم که دامنه آن "هندسه مسطحه" و برد آن از نوع عددی (integer) می‌باشد و یک ویژگی نوع شئ به نام "به وجود آمده است" را می‌سازیم که دامنه آن "توپولوژی"و برد آن "تبدیلات فضا و بعد " می‌شود.

در شکل ۲ قسمتی از روابط نشان داده شده است:

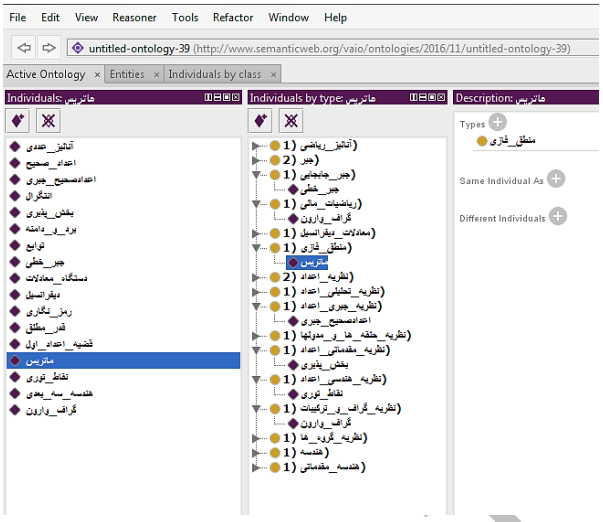
1 . Properties
2 . Object Properties
3 . Data Properties.
4. Domain
5. Range



شکل ۲. بخش ویژگی‌های کلاس‌ها و روابط

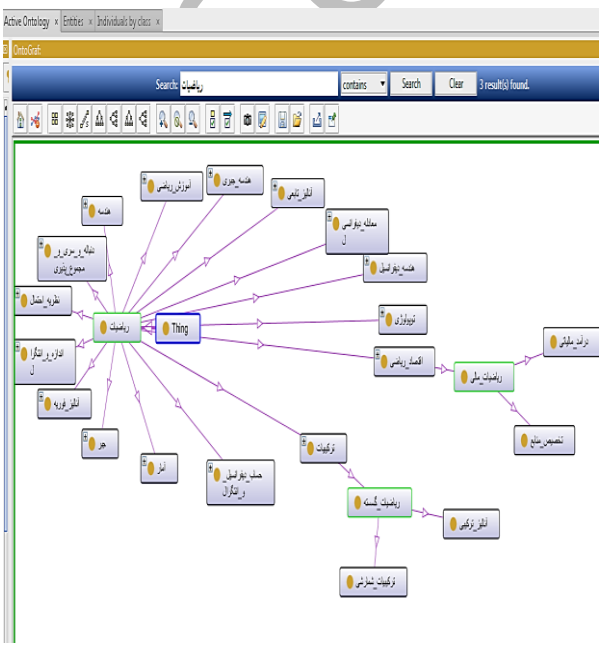
۴-۴. ایجاد نمونه‌ها:

آخرین گام ایجاد نمونه‌های منفرد در سلسله مراتب است، تعریف نمونه‌های منفرد یک کلاس مستلزم انتخاب یک کلاس، ایجاد یک نمونه منفرد آن کلاس و پر کردن ویژگی‌ها با مقادیر تعیین شده و مجاز است. در این پژوهش برای کلاس‌ها و زیر کلاس‌ها وارده نمونه‌هایی انتخاب و وارد نرم‌افزار گردید.



شکل ۳

در شکل ۴ می‌توان به صورت گراف ارتباط کلاس‌ها با یکدیگر و نمونه‌ها را مشاهده کرد این بخش از قسمت OntoGraf ایجاد می‌گردد



شکل ۴

۴. بحث و نتیجه‌گیری

در این مقاله، با مرور پژوهش‌های پیشین و با ایجاد نمونه‌ای اولیه از هستی‌شناسی ریاضی، روشی برای ایجاد هستی‌شناسی‌های حوزه ریاضی در زبان فارسی ارائه شد و با توجه به مراحل انجام شده در این پژوهش، پژوهشگران مراحل فوق را برای ایجاد هستی‌شناسی حوزه ریاضی در زبان فارسی مناسب می‌دانند. مشکل اصلی برای مراحل ساخت آنتولوژی، زمان، دقت و مزیت آن کاربردش در پایگاه‌های دانش است.

منابع

[۱] رحیمی، صالح (۱۳۹۴). نگرش‌های رایج در نمایه‌سازی و بازیابی تصاویر در محیط وب. *فصلنامه مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۲۶(۱)، ۱۳۳-۱۴۹.

[۲] زاهدی، راضیه؛ امین، غلامرضا؛ کریمی، مهرداد و علی بیک (۱۳۹۲). روش‌شناسی ایجاد هستی شناسی مبتنی بر نظام زبان واحد پزشکی مطالعه موردی: هستی‌شناسی گیاهان دارویی ایران. *فصلنامه کتابداری و اطلاع‌رسانی*. ۱۶(۳). ۸۱- ۱۰۰

[۳] آل‌احمد، ابوالفضل (۱۳۸۵). *مقدمه‌ای بر وب معنایی*. دسترس‌پذیر در www.irstu.com/?p=9984. بازیابی شده در ۱۳۹۶/۱۰/۱۵

[۴] پورخانی، محمدرضا؛ شادگار، بیتا و عصاره، علیرضا (۱۳۹۳). بررسی روش‌های انتقال پایگاه‌های داده‌های رابطه‌ای به آنتولوژی. *مطالعات کتابداری و علم اطلاعات*. ۱۳، ۵۱-۶۸.

[5] Na, J.-C.; & Neoh, H. L. (2008). Effectiveness of UMLS semantic network as a seed ontology for building a medical domain ontology. Aslib Proceedings, 60(1), 32-46.

[۶] شیخ شعاعی، فاطمه (۱۳۸۴). هستی‌شناسی و *وب معنایی. فصلنامه کتاب،۶۴(۴)، ۱۸۹-۱۹۴.*

[۷] *شمس‌فرد، مهرنوش و احمد عبدالله‌زاده بارفروش (۱۳۸۹). استخراج دانش مفهومی از متن با استفاده از الگوهای زبانی و معنایی. تازه‌های علوم شناختی. ۴(۱)، ۴۸-۶۶.*

[۸] شادگار، بیتا، علیرضا عصاره، و آزاده هراتیان نژادی .(۱۳۸۹). وب معنایی مفاهیم و تکنیک

ها. تهران: ارمغان.

[۹] پورخانی، محمدرضا؛ شادگار، بیتا و عصاره، علیرضا (۱۳۹۳). بررسی روش‌های انتقال پایگاه‌های داده‌های رابطه‌ای به آنتولوژی. *مطالعات کتابداری و علم اطلاعات*. ۱۳، ۵۱-۶۸.

[۱۰] نوروزی، مرتضی و طاهریان، محسن (۱۳۹۰). *وب معنایی. تهران: سازمان فناوری اطلاعات ایران.*

[11] Zhou X., Xu G.a and Liu L. (2011). An Approach for Ontology Construction based on Relational Database, International Journal of Research and Reviews in Artificial Intelligence, 1(1), 16-19.

[12] Ruiz-Martínez, J. M.; Valencia-García, R.; Fernández-Breis, J. T.; García Sánchez] F.; & Martínez-Béjar, R. (2011). Ontology learning from biomedical natural. language documents using UMLSExpert Systems with Applications; 38(10), 12365-12378.

[13] Bright, T. J.; Furuya, E. Y.; Kuperman, G. J.; Cimino, J. J.; & Bakken, S. (2012). Development and evaluation of an ontology for guiding appropriate antibiotic prescribing. Journal of Biomedical Informatics; 45(3), 120-128.

[۱۸] حسینی بهشتی، ملوک السادات، وفایی، سعیده و نوروزی اقبالی، مهرداد (۱۳۹۳). *اصطلاحنامه ریاضی*. تهران: چاپار.

[۱۹] زاهدی، راضیه؛ امین، غلامرضا؛ کریمی، مهرداد و علی بیک (۱۳۹۲). روش‌شناسی ایجاد هستی شناسی مبتنی بر نظام زبان واحد پزشکی مطالعه موردی: هستی‌شناسی گیاهان دارویی ایران. *فصلنامه کتابداری و اطلاع‌رسانی*. ۱۶(۳). ۸۱- ۱۰۰.

[۲۰] دوخانی، فیروزه (۱۳۹۴). چشم اندازی به نرم افزار پروتج.*گفتمان علم و فناوری* ،۱ (۷)، ۴۰۷-۴۳۲.

[۲۱] کورانت، ریچارد و رابیتز، هربرت؛ مترجم سیامک کاظمی (۱۳۹۲). ریاضیات چیست؟ تهران: نشر نی.

[14] Syamili. C, Rekha, RV. (2017). "Developing an ontology for Greek mythology", The Electronic Library, https://doi.org/.1108/EL-02-2017-0030

[۱۵] شمس‌فرد، مهرنوش (۱۳۸۱). *طراحی مدل یادگیر هستی‌شناسی: نمونه‌سازی در یک سیستم درک متن فارسی. پایان‌نامه دکتری. تهران: دانشگاه صنعتی امیرکبیر.*

[۱۶] فتحیان دستگردی، اکرم (۱۳۸۹). *مقایسه کارآمدی اصطلاحنامه و هستی‌شناسی در بازنمون دانش و بازیابی مفاهیم.* پایان‌نامه کارشناسی ارشد رشته کتابداری و اطلاع‌رسانی.

[۱۷] زاهدی، راضیه؛ امین، غلامرضا؛ کریمی، مهرداد و علی بیک (۱۳۹۲). روش‌شناسی ایجاد هستی شناسی مبتنی بر نظام زبان واحد پزشکی مطالعه موردی: هستی‌شناسی گیاهان دارویی ایران. *فصلنامه کتابداری و اطلاع‌رسانی.* ۱۶(۳). ۸۱- ۱۰۰.

[۲۲] مصاحب، غلامحسین (۱۳۵۷). تئوری مقدماتی اعداد. تهران: انتشارات سروش.

[۲۳] کاووسی، فریدون (۱۳۸۴). هندسه، زوایا، اعداد. تهران: دولتمند.

[۲۴] باندی، جان آدریان و مورتی .مترجم دارا معظمی (۱۳۸۴). *نظریه گراف وکاربردهای آن.* تهران: مرکزنشر دانشگاهی.

[۲۵] پیائوشو، هوئی (۱۳۹۵). *آنالیز فوریه کاربردی*؛ مترجم سمیه ابراهیم زاده، تهران: ماهواره.

[۲۶] پورکاظمی، محمدحسین (۱۳۹۴). *ریاضیات عمومی وکاربردهای آن.* تهران: نشر نی.

[۲۷] موحدمنش، صادق‌علی و فتحی، فائزه (۱۳۹۴). اقتصاد ریاضی. تهران: روجین مهر.

طراحی نرم‌افزار ریشه‌یابی خودکار اسامی زبان فارسی تحت وب

سمانه سلطان‌آبادی* و **محمدحسین شرف‌زاده****

چکیده

پردازش زبان‌ها یکی از اموری است مورد توجه بسیاری از پژوهشگران قرار گرفته است. بر این مبنا، هدف از انجام این پژوهش طراحی نرم‌افزار ریشه‌یابی واژگان زبان فارسی تحت وب است. ریشه‌یابی که در آن با حذف پیشوندها و پسوندها، ریشه‌ی واژه مشخص می‌شود، یکی از کاربردهای پردازش متن است. برای انجام عملیات ریشه‌یابی خودکار با رایانه، ابتدا مرز واژه‌ها در متن مشخص می‌شود تا بتوان ریشه‌ی واژه‌ها را استخراج کرد. علائم اضافی مانند ویرگول، دو نقطه، کروشه، پرانتز و ... با استفاده از فراخوانی تابع مربوط حذف می شوند. سپس ساختار کلی برنامه که شامل کلمه، طول کلمه، ریشه‌ی موقت و ریشه‌ی حقیقی می‌باشد شکل می‌گیرد. در مرحله‌ی بعد عملیات نرمال سازی در سطوح مختلف بر روی کلمات انجام می گیرد. در آخر با توجه به حروف پایانی کلمات، فراخوانی توابع مربوط و عملیات ریشه‌یابی صورت می‌پذیرد. عملیات ریشه یابی تا زمانی انجام می شود که ریشه پر نشده باشد و تا پیش از پرشدن ریشه، ریشه‌ها در یک مکان موقت به‌نام tmpRoot نگهداری می‌شوند. در این پژوهش ۴۰ تابع برای انجام عملیات ریشه یابی نوشته شده است که هر کدام از آنها برای انجام عملیات مختلفی فراخوانده می‌شوند. دیتابیسی نیز شامل ۳۵ جدول فراهم گردیده که این جداول بر اساس حروف آخر کلمات فارسی تنظیم شده‌اند. بدین ترتیب برای هر کدام از حروف دو جدول در نظر گرفته شده است. جدول دیگر، جدول بن افعال است که دربردارنده ی بن افعال ماضی، مضارع و مصادر آنها می‌باشد. این برنامه به زبان php نوشته شده است و از دیتابیس mysql برای ذخیره‌سازی جداول استفاده شده است.

واژه‌های کلیدی: ریشه‌یابی خودکار، زبان‌شناسی رایانه‌ای، پردازش زبان‌های طبیعی، نرم‌افزار، زبان برنامه‌نویسی php.

---

* کارشناس ارشد زبان‌شناسی همگانی، واحد مرودشت، دانشگاه آزاد اسلامی، مرودشت، ایران

** استادیار گروه زبان‌شناسی، واحد مرودشت، دانشگاه آزاد اسلامی، مرودشت، ایران (*نویسنده مسؤول)

۱. مقدمه

امروزه با گسترش کاربرد زبان در سیستم‌های رایانه‌ای، نیاز به پردازش متون در این سیستم‌ها، بیش از پیش احساس می‌شود. ریشه‌یابی یکی از کاربردهای پردازش متن است. پردازش متن شامل چهار سطح است: پردازش لغوی، پردازش ساختواژی، پردازش نحوی و پردازش معنایی [۱].

در زبان‌های طبیعی از بین کل واژگان می‌توان تعداد معدودی واژه یافت که واژه‌های ریشه در نظر گرفته می‌شوند، و بقیه واژگان از این ریشه‌ها اشتقاق می‌شوند. به عبارت دیگر هر واژه دارای ریشه‌ای است که ایده و معنا و مفهوم اصلی آن واژه را در بر می‌گیرد. اشتقاق یک واژه از ریشه اصلی سبب می‌شود تا مفهوم اصلی واژه شکل بهتری به خود بگیرد و یا واژه برای بروز نقش نحویِ خود در جمله آماده شود. در زبان فارسی (و خانواده زبان‌های هند و اروپایی) عملیات اشتقاق و ساخت واژگان با ترکیب واژه‌های ریشه و الحاق پسوند و پیشوندهای مختلف صورت می‌گیرد. هدف از ریشه‌یابی، زدودن الحاقات و یافتن جوهره اصلی واژه است. هر چند در واقعیت گاهی الحاقات واژه معنای آن را چنان تغییر می‌دهند که حذف آنها موجب از بین رفتن معنای اصلی می‌شود. ریشه‌یابی به فرایندی اطلاق می‌شود که در آن با حذف پیشوندها و پسوندها نهایتاً ریشه‌ی یک واژه مشخص می‌شود. به عبارت دیگر بعد از انجام عملیات ریشه‌یابی، واژه‌ی حاصل غیر قابل تجزیه خواهد بود. چنانچه این عملیات به کمک کامپیوتر انجام شود به آن ریشه‌یابی خودکار گفته می‌شود [۲].

پردازش لغوی در عملیات ریشه‌یابی به طور گسترده مورد استفاده قرار می‌گیرد که از آن جمله، می‌توان به تعیین مرز واژه‌ها و جمله‌ها، یک‌دست‌سازی پیکره‌ی متنی، شناسایی حروف اضافه اشاره کرد.

کاربرد عمده‌ی ریشه‌یابی واژه‌ها در خلاصه‌سازی متن است که در مرحله‌ی پیش‌پردازش انجام می‌گیرد. ریشه‌یابی کردن واژه‌ها تضمین می‌کند که سندهایی که همگی شامل اشتقاق‌های متفاوتی از کلمه‌ی موجود در پرس و جو هستند، در مجموعه جواب نهایی وجود دارند.

از دیگر سو، عملیات ریشه‌یابی، در ریشه‌یابی کلمات کلیدی آیتم‌ها در سیستم‌های توصیه‌گر بسیار کارا است. سیستم‌های توصیه‌گر سیستم‌هایی هستند که با دادن پیشنهادات

مناسب با توجه به علایق یک کاربر و تحلیل رفتار او بر اساس پیشینه‌ی عملکردش، وی را از صرف وقت در مرور تمام آیتم‌ها باز می‌دارد و اقدام به پیشنهاد مناسب‌ترین اقلام به وی می‌نماید. در نتیجه یکی از قسمت‌های اصلی این سیستم‌ها، استخراج مفاهیم از آیتم‌ها و رفتار کاربران می‌باشد که این کار با تکنیک‌های پردازش زبان طبیعی مانند ریشه‌یابی کلمات کلیدی آیتم‌ها میسّر می‌باشد [۳].

از دیگر کاربردهای ریشه‌یابی می‌توان به افزایش کارایی سیستم‌های بازیابی اطلاعات اشاره کرد. یکی از مهمترین موضوعات در پردازش زبان طبیعی و بازیابی اطلاعات، یافتن ریشه‌ی کلمات می‌باشد. ریشه‌ی کلمه جزئی از کلمه است که پس از حذف وندهای کلمه (پیشوند، پسوند و میانوند) باقی می‌ماند. یکی از روش‌های افزایش کارایی سیستم‌های بازیابی اطلاعات استفاده از ریشه‌یابی کلمات است؛ زیرا اشتقاقات مختلف یک کلمه به ریشه‌ی آن کلمه تبدیل می‌شوند؛ در نتیجه جستجو بر اساس ریشه‌ی کلمه انجام خواهد شد و اندازه ساختار ایندکس کاهش می‌یابد[۴].

از دیگر کاربردهای ریشه‌یابی، می‌توان به سیستم‌های غلط‌یاب املایی اشاره کرد. غلط‌یاب املایی فارسی یکی از ابزارهای مهمی است که در راستای کمک به نویسنده‌ی یک متن فارسی می‌تواند به او کمک شایانی در یافتن و درست کردن غلط فارسیِ نوشته شده در یک متن نماید. یک غلط‌یاب املایی شامل سه مرحله است: ۱- شناسایی خطا ۲- ساخت پیشنهادهایی برای تصحیح ۳- رتبه بندی پیشنهادها. در غلط‌یاب املایی برای ایجاد پیشنهادهایی برای تصحیح خطا از یک لغت‌نامه استفاده می‌شود؛ هرچه حجم کلمات لغت‌نامه بیشتر باشد زمان اجرای عملیات غلط‌یابی افزایش می‌یابد. اگر عملیات ریشه‌یابی قبل از غلط‌یابی بر روی کلمه‌ی ورودی انجام شود، حالات مختلفی که با این ریشه‌ی پیشنهادی به کاربر داده می‌شود دقیق‌تر و در مدت زمان کمتری انجام می‌شود. بنابراین یک ریشه‌یاب کارآمد می‌تواند در بهبود کارایی غلط‌یاب املایی بسیار موثر باشد [۵]. همانطور که گفته شد پردازش متن شامل چهار سطح است: پردازش لغوی، پردازش ساختواژی، پردازش نحوی و پردازش معنایی. هر یک از کاربردهای فراوان پردازش متن، از جمله بازیابی اطلاعات، ریشه‌یابی واژه‌ها، خلاصه‌سازی متن، خطایابی املایی، درک، تولید، ترجمه، پرسش و پاسخ، استخراج دانش از متون و موارد دیگر با توجه به گستردگی و پیچیدگی، در یک یا چند سطح فوق به انجام می‌رسد [۶].

پردازش زبان‌های طبیعی به کمک رایانه، با اشکالات و چالش‌های بسیاری نیز همراه می‌باشد. ویژگی‌های خاص خط و زبان فارسی موجب شده تا برخی چالش‌های خاص پیرامون پردازش زبان فارسی به وجود آیند که در زبان های دیگر اصلاً مطرح نیستند. یکی از این اشکالات، نبودن دستور خط جامعی متناسب با نیازهای سامانه‌های پردازش متون است. فرهنگستان زبان و ادب فارسی از سال ۱۳۷۲ شروع به بررسی، گردآوری و تدوین دستورِ خط فارسی نمود، اما هدف اصلی از نگارش آن، یکسان سازی چهره‌ی خط جهت کاربرد در رایانه نبود و در بسیاری از موارد، دست نگارندگان برای انتخاب شکل نوشتار واژه باز گذاشته شد که نتیجه‌ی آن چیزی جز پیدایش ابهام در تشخیصِ رایانه‌ای واژه‌ها نیست. پردازش واژگانیِ کلیه‌ی زبان‌های طبیعی امری دشوار است. ترکیب واژگان، منجر به تشکیل واژگانی می‌شود که ممکن است در اثر بی‌دقتیِ کاربران، از دید رایانه به دو یا چند شکل مختلف خوانده شوند. زبان فارسی علاوه بر فاصله گذاری معمول در دیگر زبان‌ها، فاصله‌ی درون واژه‌ای[1] نیز دارد که از قوانین مشخص و دقیقی پیرامون نحوه‌ی فاصله‌گذاری پیروی نمی‌کند؛ مثلا در جایی که منظور نویسنده "می خورده است، اگر در اثر بی‌دقتی می‌خورده است نوشته شود، رایانه قادر به تشخیص واژه‌ی اصلی نخواهد بود [۷]. در این اثر سعی شده تا ابعاد، پیچیدگی‌ها و چالش‌های پردازشی زبان فارسی، خصوصاً با رویکرد ریشه‌یابی اسامی زبان فارسی مورد بررسی قرار گیرد، راهکارهای مواجهه و مرتفع ساختن برخی از این چالش‌ها مطرح شده، در نهایت الگوریتمی برای ریشه‌یابی خودکار واژگان ارائه شود.

۶-۱-۱- ۲ انواع الگوریتم‌های ریشه‌یابی در زبان فارسی

به‌طور کلی عملیات ریشه‌یابی کلمات در دو دسته‌ی ساختاری و غیر ساختاری قرار می‌گیرد.

۲-۱ الگوریتم‌های غیر ساختاری

۲-۱-۱ الگوریتم ریشه‌یاب آماری

ریشه‌یاب آماری که در آن از روشی مبتنی بر گراف و مدل آماری استفاده شده است

---

1 نشان نویسه‌ی، صحیح که می‌شود تعبیر(Pseudo-space)فاصله شبه یا فاصله نیم به واژه‌ای درون فاصله‌ی این معمولاً است U+200C با کدZero Width Non-Joiner (ZWNJ)کد یونی کد گذاری در آن دهنده‌ی.

نمونه‌ای از الگوریتم‌های غیر ساختاری است. در روش آماری یک مجموعه از کلمات زبان در نظر گرفته شده و هر کلمه به دو زیر رشته تقسیم می‌شود. زیر رشته‌ی اول پیشوند و زیر رشته‌ی دوم پسوند در نظر گرفته می‌شود. سپس هر زیر رشته به عنوان یک گره از گراف در نظر گرفته شده و یک یال بین دو گره نشان دهنده‌ی این است که از ترکیب این دو زیر رشته، یک کلمه از مجموعه‌ی لغات بدست می‌آید. برای نشان دادن این تأثیر متقابل، از یک نمادگذاری استفاده شده است تا پیشوند بهینه که همان ریشه است و پسوند بهینه که همان اشتقاق است پیدا شود. در تشخیص ریشه و اشتقاق‌ها هم از شمردن تعداد تکرار آنها در حالات مختلف تقسیم در کل کلمات و هم از یک تحلیل آماری بر روی لینک‌های بین زیر رشته‌ها استفاده می‌شود [۸]. شایان ذکر است این روش نیازمند هیچ قانون زبان‌شناسی نیست و به صورت مستقل از قوانین زبان‌شناسی عمل می‌کند. مهمترین مشکل این روش پیاده‌سازی آن می‌باشد.

۲-۱-۲ الگوریتم باچین[1]

الگوریتم باچین یک ریشه‌یاب آماری است که در آن از روشی مبتنی بر گراف و مدل آماری استفاده شده است و نمونه‌ای از الگوریتم‌های غیر ساختاری است.

۲-۲- الگوریتم‌های ساختاری

۲-۲-۱ الگوریتم بن

این ریشه‌یاب از تعدادی قانون تشکیل شده است که بر اساس آنها طولانی‌ترین رشته‌ی پیشوند یا پسوند را که در کلمه هست، حذف می‌کند تا به ریشه‌ی کلمه برسد. در اکثر موارد ریشه‌ی به دست آمده ریشه‌ی حقیقی است؛ ولی بعضی مواقع این ریشه‌یاب‌ها برای رسیدن به ریشه‌ی حقیقی مجبور به تغییر واژه‌ی به دست آمده می‌شوند؛ برای این کار دو روش کدگذاری مجدد و تطبیق جزئی وجود دارد که الگوریتم بن از روش کدگذاری مجدد استفاده می‌کند. روش کدگذاری مجدد یک تغییر وابسته به زمینه به شمار می‌رود. در این تغییر با توجه به دنباله‌ی حروفی که در واژه‌ی حاصله وجود دارد، طبق قواعدی بعضی از حروف واژه تبدیل به

1 Bacchin

حروف دیگری می‌شوند. به طور نمونه در قاعده AxC→AyC که A و C زمینه را نشان می‌دهند، دنباله حروف x به دنباله حروف y تبدیل می‌شود [۲].

۲-۲-۲ الگوریتم کاظم تقوا

این الگوریتم زیر رشته‌ای از کلمه‌ی ورودی را که در لیست انواع پسوندهای متداول فارسی موجود است، پیدا می‌کند. اگر چندین پسوند با کلمه مطابقت پسوندی داشته باشد، طولانی‌ترین پسوندی که با حذف آن پسوند، ریشه‌ی باقیمانده حداقل سه حرف داشته باشد، انتخاب می‌گردد [۴].

به دلیل اینکه این ریشه‌یاب از طول رشته برای تعریف کران پائین محتوای ریشه استفاده می‌کند، (در حال حاضر مینیمم طول ریشه ۳ است) در بعضی موارد باعث خطا می‌گردد بخصوص زمانی‌که یک زیررشته که قسمتی از یک کلمه کوتاه است، به اشتباه به عنوان یک پسوند در نظر گرفته شود [۹].

۲-۲-۳ الگوریتم ریشه‌یاب جدولی[1]

در روش ریشه‌یاب جدولی ریشه‌ی هر واژه در یک جدول نگهداری می‌شود. در این روش با جستجوی واژه در این جدول، ریشه‌ی واژه مشخص می‌گردد. اگرچه این روش نتایج خوبی دارد؛ اما این روش مشکلاتی نیز به همراه دارد - که در مورد زبان فارسی شامل موارد زیر است:- اولاً برای واژه‌های فارسی این اطلاعات فعلاً وجود ندارد و ثانیاً نگهداری این جدول سربار زیادی برای سیستم خواهد داشت و سرعت سیستم بطور قابل ملاحظه‌ای کاهش می‌یابد [۲].

۲-۲-۴ روش n-gram

روش n-gram کلمات را بر اساس تعداد دیاگرام‌ها و چند گرام‌های مشترک آنها تلفیق می‌کند. همچنین کلمات و ریشه‌های مربوطه در یک جدول نگهداری می‌شوند؛ سپس عملیات ریشه‌یابی بوسیله‌ی جستجو در جدول انجام می‌پذیرد.

---

1 Table lookup

۲-۲-۵ الگوریتم گوناگونی پسین[1]

روش گوناگونی پسین از بسامدهای توالی حروف در متن به عنوان اساس ریشه‌یابی بهره می‌جوید [۱۰]. شایان ذکر است که در دو روش گوناگونی پسین و n-gram ریشه‌ی حقیقی تولید نمی‌شود.

۲-۲-۶ الگوریتم پایین به بالا

این الگوریتم نیز از روش ساختاری بهره می‌جوید و به صورت پایین به بالا عمل می‌کند. این ریشه‌یاب از سه بخش تشکیل شده است. ۱- برجسب‌گذاری زیر رشته‌ها ۲- تطبیق قوانین ۳- تطبیق ضد قوانین [۲].

۲-۲-۷ الگوریتم کراوتز بهبود یافته

در سال ۲۰۰۶ توسط رضا حسامی فرد و غلامرضا قاسم ثانی الگوریتمی بهبود یافته از الگوریتم کراوتز ارائه شد. در این الگوریتم از نقش کلمات علاوه بر وجود یا عدم وجود آنها در فرهنگ لغت استفاده می‌شود.

به عنوان مثال در فرآیند ریشه‌یابی کلمه‌ی "being" ابتدا "bee" جستجو می‌شود که در فرهنگ لغت هم موجود می‌باشد. بنابراین به عنوان ریشه برگردانده می‌شود اما با توجه به اینکه bee یک اسم است نمی‌توان به آن پسوند ing افزود. بنابراین الگوریتم این ریشه را به عنوان ریشه‌ی اصلی انتخاب نمی‌کند و به این ترتیب با ادامه‌ی الگوریتم، ریشه‌ی درست کلمه که همان "be" است برگردانده می‌شود [۹].

۲-۲-۸ الگوریتم لاوینز

این الگوریتم شامل ۲۵۰ پسوند است که طولانی‌ترین پسوند متصل به کلمه را حذف می‌کند؛ با این شرط که کلمهٔ باقی‌مانده حداکثر سه نویسه داشته باشد [۱۱].

---

1 Successor variety

۲-۲-۹ الگوریتم پورتر

ریشه‌یاب پورتر ریشه‌یاب کاهش‌دهنده‌ی ادغامی و از نوع ریشه‌یاب ساختاری برای زبان انگلیسی است که توسط مارتین پورتر در دانشگاه کمبریج در سال ۱۹۸۰ ارائه شد [۹].

در این روش با حذف پسوندهای کلمات، از تعداد واژه‌های منحصربه‌فرد در بازیابی اطّلاعات کاسته می‌شود. درنتیجه موجب بالا رفتن کارایی سیستم خواهد شد [۱۱].

در زبان انگلیسی با توجه به قدمت استفاده از گرامر فرمال، قواعد خاصی استخراج گردیده است که تا حد زیادی کار ریشه‌یابی را انجام می‌دهد. الگوریتم پورتر با در نظر گرفتن این قواعد کار ریشه‌یابی را انجام می‌دهد. این الگوریتم معروف‌ترین و متداول‌ترین الگوریتم در زبان انگلیسی به شمار می‌آید [۸].

۲-۲-۱۰ الگوریتم کراوتز

الگوریتم کراوتز برای اولین بار در سال ۱۹۹۳ در یک مقاله توسط کراوتز معرفی شد. این الگوریتم از روش‌های ساخت‌واژه‌ای و از یک فرهنگ لغت برای آزمودن ریشه‌های یافت شده استفاده می‌کند. این الگوریتم برای زبان‌هایی که ساخت کلمات در آنها قانونمند است، کارائی خوبی را نشان داده است و در ماشینهای مترجم می‌تواند بکار گرفته شود [۹].

۳ الگوریتم ریشه‌یابی پژوهش

در این بخش روند انجام الگوریتم پیشنهادی به صورت کامل شرح داده خواهد شد. همانطور که قبلاً ذکر شده مبنای الگوریتم، بر اساس حذف پسوند و پیشوند با توجه به حروف پایانی کلمات، بطور بازگشتی می‌باشد. بدین صورت که روش حذف پسوند تا زمانی که کلمه به ریشه‌ی حقیقی برسد ادامه می‌یابد. در این میان اگر کلمه دارای پیشوند باشد، پیشوند نیز حذف می‌شود.

الگوریتم پیشنهادی در این پژوهش، از یک روش ترکیبی که تلفیقی از روش حذف پیشوند و پسوند (الگوریتم بن)، روش جدولی و نیز پاره‌ای از قوانین موجود است، استفاده می‌شود. اولین قدم برای ریشه‌یابی واژه‌های یک متن، جدا کردن آنها از یکدیگر است؛ برای این منظور تابع WordSplit فراخوانده می‌شوند و کلمات در درون یک جمله تشخیص داده

می‌شوند؛ به عبارت دیگر باید مرز واژه‌ها در متن دقیقاً معین باشد تا بتوانیم ریشه‌ی هر واژه را استخراج نماییم. در اکثر زبان‌های طبیعی شناسایی واژه‌ها از طریق بررسی علایم قابل انجام است. این علایم عبارتند از: فضای خالی، ".", ">", "<", "[", "]", "-", "_", "/" و ... .

در مرحله‌ی بعد ساختار کلی برنامه که شامل کلمه، طول کلمه، ریشه‌ی موقت و ریشه‌ی حقیقی توسط تابع Structuring شکل می‌گیرد.

سومین مرحله برای انجام عملیات ریشه‌یابی یک‌دست‌سازی پیکره‌ی متنی است که توسط تابع نرمال‌سازی در سطوح مختلف بر روی کلمات انجام می‌گیرد. گاهی دو واژه‌ی یکسان با املاهای متفاوت به عنوان دو واژه‌ی مختلف در نظر گرفته می‌شوند. به عنوان مثال واژه‌ی درخت‌ها را می‌توان به سه صورت زیر نوشت؛

بدون فاصله	با فاصله	نیم‌فاصله
درختها	درخت ها	درخت‌ها

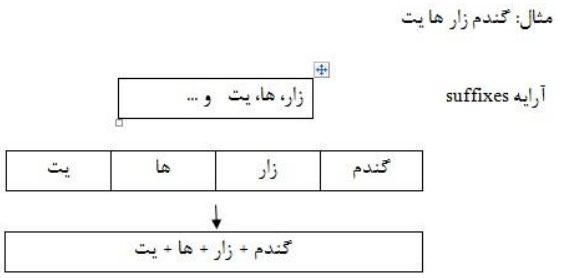
در این مرحله تابع Normalize فراخوانی می‌شود. نرمال‌سازی شامل سه مرحله می‌باشد: ۱- نرمال‌سازی پیشوند، ۲- نرمال‌سازی پسوند ۳- نرمال‌سازی خود کلمه. در حالت اول اگر کلمه پیشوند باشد به صورت موقت در یک متغیر نگاه داشته می‌شود که زمان اضافه شدن کلمه‌ی بعدی در آرایه‌ی Normalize به کلمه‌ی نرمال شده اضافه شود. حالت دوم اگر کلمه پسوند باشد به آخرین کلمه‌ای که نرمال شده در آرایه‌ی نرمال‌سازی اضافه می‌شود. در حالت سوم کلمه به آرایه‌ی نرمال‌سازی می‌رود که بعداً ریشه‌یابی شود. شایان ذکر است تابع Normalize تا چندین سطح، کلمات را نرمال می‌کند و قادر به تشخیص فاصله و نیم‌فاصله می‌باشد. به عنوان مثال کلمه‌ی "می رود" را به "می‌رود" تبدیل می‌کند و نیز "نمی خواهم بروم" را به "نمی‌خواهم بروم" تبدیل و سپس عملیات ریشه‌یابی انجام می‌شود.

در آخر با توجه به حروف پایانی کلمات، توابع مربوطه فراخوانی و عملیات ریشه‌یابی صورت می‌پذیرد. شایان ذکر است عملیات ریشه‌یابی تا زمانی انجام می‌شود که root پر نشده باشد و تا پیش از پر شدن root، ریشه‌ها در یک مکان موقت به نام tmpRoot نگهداری می‌شوند.

در آخر تابع RemoveSuffix فراخوانی می‌شود که بر اساس حروف پایانی کلمات، عملیات حذف پسوند انجام می‌شود. به عنوان مثال اگر کلمه به "ا" ختم شده باشد به تابع CheckSuffix_alef رفته و پردازش لازم جهت ریشه‌یابی بر روی کلمه صورت می‌گیرد. این مراحل تا CheckSuffix_ y ادامه می‌یابد. لازم به ذکر است اگر کلمه دارای پیشوند باشد تابع RemovePrefix فراخوانی شده و عملیات حذف پیشوند نیز انجام می‌گیرد.

۶-۲- ۳-۱ تابع نرمال‌سازی[1]

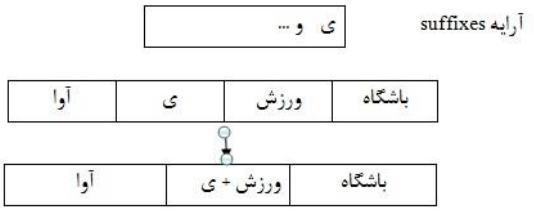
ورودی این تابع، آرایه‌ای از کلمات است؛ به عنوان مثال، کلمه‌ی «گندم زارها یت». در این مثال، آرایه شامل چهار خانه است. خانه‌های آرایه با فاصله (space) قابل تشخیص هستند. همچنین تمامی پسوندها در آرایه‌ای به نام sufffixes قرار می‌گیرد. بنابراین «زار»، «ها»، و «یت» در آرایه‌ی sufffixes قرار می‌گیرند. در یک حلقه تا زمانی که به انتهای کلمات برسیم، اگر کلمه‌ی در حال بررسی (گندم) در آرایه‌ی sufffixes نباشد، آن کلمه در آرایه‌ای به نام normalized قرار می‌گیرد و در غیر این صورت برای بررسی کلمه‌ی زار، ابتدا ریشه‌ی پسوندها را یک مقدار پیش‌فرض ثابت در نظر می‌گیریم (به عنوان مثال ریشه‌ی زار را x در نظر می‌گیریم). اگر ریشه‌ی کلمه‌ی قبل از خانه‌ی آرایه‌ی در حال بررسی برابر با این مقدار پیش‌فرض ثابت نباشد (به عبارت ساده‌تر اگر کلمه‌ی قبل از کلمه‌ی در حال بررسی، پسوند نباشد، که در این مثال این گونه است)، کلمه به عنوان ریشه‌ی نرمال شده در نظر گرفته می‌شود و در غیر این صورت کلمه به آخرین کلمه‌ی آرایه‌ی نرمال‌سازی اضافه می‌شود. این روند تا پایانِ بررسی تمامی کلمات آرایه، ادامه می‌یابد.



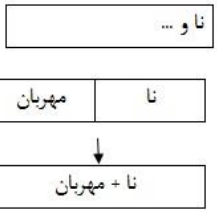
1 Normalized

برای نرمال‌سازی پیشوند، یک آرایه به نام prefixes در نظر گرفته شده است کـه لیسـت پیشوندها در آن قرار می‌گیرد. اگر کلمه‌ی در حال بررسی در آرایـه‌ی prefixes نباشـد، آنگـاه آن کلمه به آرایه‌ی نرمال‌سازی، اضـافه مـی‌شـود. اگـر کلمـه‌ی در حـال بررسـی در آرایـه‌ی prefixes باشد، آنگاه آن کلمه به ابتدای آرایه‌ی کلمات، اضافه می‌شود و سپس عملیات ریشـه یابی روی آن صورت می

مثال: باشگاه ورزشی آوا



مثال: نامهربان



در این بخش برای نمونه مراحل حذف پسوند و پیشوند در کلمات مختوم به الـف را شـرح داده و قابل ذکر است این مراحل برای حروف "ب" تا "ی" نیز انجام می‌شود.

۳-۲ پسوندهای مختوم به الف شامل: الف -ها - وا - نا – یا می‌باشد

برای حذف پسوند در کلماتی که به الف ختم می‌شوند به موارد زیر برمی‌خوریم:

- الف جزء خود کلمه است؛ مانند: ابتدا. در کلمه‌ای مانند گرما، ریشـه‌ی کلمـه "گـرم" است و "ا" باید حذف شود و یا در کلمه‌ای مانند ملکا، ریشـه‌ی کلمـه "ملـک" اسـت و "ا" بایـد حذف شود. در کلمه‌ای مانند گفتا، ریشه‌ی کلمه، فعل "گفت" است و "ا" باید حذف شود.
- کلمه مختوم به "ها" باشد؛ مانند درخت‌ها که "ها" در اینجا پسوند است و باید حذف شود. در کلمه‌ای مانند انتها "ها" جزء خود کلمه است و نبایـد حـذف شـود. از طـرف دیگـر در

کلمه‌ای مانند رها، ریشه‌ی کلمه "ره" است و "ا" باید حذف شود. در این میان کلمه‌ای مانند فقها جمع مکسر است و ریشه‌ی آن فقیه است.

- کلمه مختوم به "وا" باشد؛ مانند نانوا که در این صورت "وا" پسوند است و باید حذف شود و نان به عنوان ریشه برگردانده شود. در کلمه‌ای مانند تقوا "وا" پسوند نیست و نباید حذف شود. در کلمه‌ی شنوا، بن فعل یعنی شنو ریشه است و "ا" باید حذف شود.
- کلمه مختوم به "نا" باشد؛ مانند تنگنا که در این صورت "نا" پسوند است و باید حذف شود و تنگ به عنوان ریشه برگردانده شود. این در حالی است که در کلمه‌ی آشنا، "نا" جزء خود کلمه است و نباید حذف شود و یا در کلمه‌ای مانند روشنا "ا" باید حذف شود و روشن به عنوان ریشه برگردانده شود. در کلمه‌ای مانند توانا، ریشه‌ی کلمه، توان است و "ا" باید حذف شود. بنابراین الگوریتم باید به گونه‌ای نوشته شود که این موارد را در نظر بگیرد.
- کلمه مختوم به "یا" باشد؛ مانند خدایا و جویا که "یا" پسوند است و ریشه‌ی کلمه خدا و جو است. در کلمه‌ی ساقیا، ساقی ریشه است و "ا" باید حذف شود. در کلمه‌ی اشیا "یا" جزء خود کلمه است و نباید حذف شود.

۳-۳ مراحل حذف پسوند و پیشوند در کلمات مختوم به الف

ابتدا جدول استثنائات چک می‌شود، کلماتی که خارج از الگوی الگوریتم پیشنهادی است، در جدول استثنائات چک می‌شوند. قابل ذکر است که جدول استثنائات از الگوی خاصی پیروی نمی‌کند؛ به عنوان مثال، در مرحله‌ی حذف پسوند در کلماتی که به الف ختم می‌شوند، چون تعداد کلماتی که با "ها" جمع بسته می‌شوند زیاد است، کلماتی که در آنها "ها" جزء خود کلمه است در جدول استثنائات برده می‌شوند؛ لذا اگر کلمه در جدول استثنائات بود به عنوان ریشه برگردانده می‌شود.

۳-۳-۱ کلمه‌های مختوم به "ها"

اگر کلمه‌ای به "ها" ختم شده باشد ابتدا جدول استثنائات چک می‌شود؛ اگر کلمه در جدول استثنائات باشد به عنوان ریشه‌ی حقیقی برگردانده می‌شود(تنها --> تنها).

اکنون اگر کلمه دارای پیشوند باشد تابع RemovePrefix صدا زده می‌شود و پیشوند در

این مرحله حذف شده و کلمه در جدول اسم جستجو می‌شود؛ اگر کلمه در جدول اسم باشد به عنوان ریشه‌ی حقیقی برگردانده می‌شود (بی‌اشتها --> اشتها).

در غیر این صورت "ها" از انتهای کلمه حذف می‌شود و به عنوان ریشه‌ی موقت در این مرحله برگردانده می‌شود. به عنوان مثال در کلمه‌ی زیبایی ها، "ها" حذف شده و زیبایی به عنوان ریشه‌ی موقت در نظر گرفته می‌شود؛ سپس کلمه‌ی زیبایی به تابعِ حذف پسوند در کلماتی که به "ی" ختم شده رفته و در آنجا عملیات خذف پسوند بر روی کلمه انجام شده و زیبا به عنوان ریشه‌ی موقت برگردانده می‌شود(زیبایی‌ها --> زیبا). سپس کلمه‌ی زیبا به تابعِ حذف پسوند در کلماتی که به "الف" ختم شده رفته و در آنجا عملیات خذف پسوند بر روی کلمه انجام شده و زیب به عنوان ریشه‌ی حقیقی برگردانده می‌شود(زیبا --> زیب).

کلماتی مانند رها که به "ها" ختم شده‌اند ولی "ه" جزء خود کلمه است و "ا" باید خذف شود چون تعدادشان کم است به جدول استثنائات برده می‌شوند و در آنجا چک می‌شوند(رها --> ره).

همچنین کلماتی مانند فقها که جمع مکسرند به جدول جمع مکسر برده شده و در آنجا چک می‌شوند(فقها --> فقیه).

۳-۳-۲ کلمه‌های مختوم به "وا"

اگر کلمه‌ای به "وا" ختم شده باشد ابتدا جدول استثنائات چک می‌شود اگر کلمه در جدول استثنائات باشد به عنوان ریشه‌ی حقیقی برگردانده می‌شود (شنوا --> شنو).

بعد از چک کردن جدول استثنائات "وا" از انتهای کلمه خذف شده و سپس کلمه در جدول اسم‌ها جستجو می‌شود؛ اگر در جدول بود به عنوان ریشه‌ی حقیقی برگردانده می‌شود (نانوا --> نان).

در غیر این صورت اگر کلمه دارای پیشوند باشد تابع RemovePrefix صدا زده می‌شود و پیشوند در این مرحله حذف شده و کلمه به عنوان ریشه‌ی حقیقی برگردانده می‌شود(بی‌پروا --> پروا).

۳-۳-۳ کلمه‌های مختوم به "نا"

اگر کلمه‌ای به "نا" ختم شده باشد جدول استثنائات چک می‌شود؛ اگر کلمه در جدول استثنائات باشد به عنوان ریشه‌ی حقیقی برگردانده می‌شود(رسانا --> رسانا).

در غیر این صورت برای کلماتی مانند تنگنا که به "نا" ختم شده‌اند و "نا" پسوند است، "نا" حذف شده و کلمه در جدول اسم‌ها جستجو می‌شود اگر در جدول بود به عنوان ریشه‌ی حقیقی برگردانده می‌شود(تنگنا --> تنگ).

در غیر اینصورت برای کلماتی مانند روشنا که به "نا" ختم شده‌اند ولی "نا" پسوند نیست و "ا" پسوند است، "ا" حذف می‌شود سپس کلمه در جدول اسم‌ها جستجو می‌شود اگر در جدول بود به عنوان ریشه‌ی حقیقی برگردانده می‌شود (روشنا --> روشن).

در غیر این صورت برای کلماتی مانند مانا که در آن ریشه‌ی کلمه بن فعل می‌باشد "ا" حذف شده و کلمه در بن فعل جستجو می‌شود. اگر کلمه در جدول بن فعل وجود داشت به عنوان ریشه برگردانده می‌شود؛ در این مرحله اگر کلمه دارای پیشوند باشد تابع RemovePrefix صدا زده می‌شود و پیشوند نیز در این مرحله حذف می‌شود (نابینا --> بین).

در غیر این صورت خود کلمه ریشه است. شایان ذکر است اگر کلمه دارای پیشوند باشد تابع RemovePrefix صدا زده می‌شود و پیشوند نیز در این مرحله حذف می‌شود (بامعنا --> معنا).

۴-۳-۳ کلمه‌های مختوم به "یا"

اگر کلمه‌ای به "یا" ختم شده باشد ابتدا جدول استثنائات چک می‌شود. اگر کلمه در جدول استثنائات باشد به عنوان ریشه‌ی حقیقی برگردانده می‌شود (سویا --> سویا). همانطور که قبلاً ذکر شد، کلمات جمع مکسر در جدول استثنائات چک می‌شود (سجایا --> سجیه).
در غیر این صورت اگر حرف قبل از "ی"، "ا" یا "و" باشد، ("ی" در اینجا میانجی است) "یا" حذف می‌شود و باقیمانده‌ی کلمه به عنوان ریشه‌ی حقیقی برگردانده می‌شود (جویا --> جو)، (پایا --> پا).

برای کلماتی مانند ساقیا که به "یا" ختم شده‌اند ولی "یا" پسوند نیست و "ا" پسوند است،

"ا" حذف می‌شود سپس کلمه در جدول اسم‌ها جستجو می‌شود اگر در جدول بود به عنوان ریشه‌ی حقیقی برگردانده می‌شود(ساقیا --> ساقی).

در غیر این صورت خود کلمه ریشه است. در این مرحله اگر کلمه دارای پیشوند باشد تابع RemovePrefix صدا زده می‌شود و پیشوند نیز در این مرحله حذف می‌شود (بی‌حیا --> حیا).

۳-۳-۵ حالت پیش‌فرض

در حالت آخر اگر برای کلماتی که به "ها"،"نا"، "وا"، "یا"، "تا" ختم نشده باشد، مانند کلمه‌ی گرما که "ا" باید از انتهای کلمه حذف شود و گرم به عنوان ریشه برگردانده شود "ا" راحذف کرده و سپس کلمه در جدول اسم‌ها جستجو می‌شود. اگر در جدول بود به عنوان ریشه‌ی حقیقی برگردانده می‌شود؛ قابل ذکر است اگر کلمه دارای پیشوند باشد، تابع RemovePrefix صدا زده می‌شود و پیشوند نیز در این مرحله حذف می‌شود (ناشکیبا --> شکیب).

برای کلماتی مانند "دارا" کلمه در جدول فعل‌های مضارع جستجو می‌شود اگر در جدول بود به عنوان ریشه‌ی حقیقی برگردانده می‌شود.

برای جلوگیری از حذف کلماتی که با "با"، "بی"، "تا"، "ب"، شروع می‌شوند آنها را در جدول اسم قرار می‌دهیم، مانند کلمه‌ی بیابان‌زدا که در جدول اسم‌ها جستجو می‌شود اگر در جدول بود به عنوان ریشه‌ی حقیقی برگردانده می‌شود؛ (بیابان‌زدا --> بیابان‌زدا).

در غیر این صورت کلمه ریشه است. همچنین اگر کلمه دارای پیشوند باشد تابع RemovePrefix صدا زده می‌شود و پیشوند نیز در این مرحله حذف می‌شود (بی‌وفا --> وفا).

۴ ارزیابی ریشه‌یاب

یکی از روش‌های ارزیابی ریشه‌یاب‌ها، بررسی میزان صحت ریشه‌های حاصل از ریشه‌یاب است. به منظور ارزیابی این ریشه‌یاب، فهرستی از کلمات برای هر یک از حروف، به صورت تصادفی انتخاب گردیده است و به عنوان ورودی به ریشه‌یاب داده شده است که میزان صحت

ریشه‌های به دست آمده از ریشه‌یابِ پیشنهادی در جدول زیر آمده است.

تعداد واژه‌ها	تعداد ریشه‌های صحیح	درصد ریشه‌های صحیح
۲۷۹۰	۲۷۶۹	۹۹/۲۴

۵ نتیجه‌گیری

هدف از انجام این پژوهش، طراحی نرم‌افزار ریشه‌یابی خودکار اسامی زبان فارسی تحت وب است. برای آشنایی بیشتر با این نرم‌افزار، به آدرس www.samasoltanistemmer.ir مراجعه فرمایید.

ریشه‌یابی یکی از کاربردهای پردازش متن است. پردازش متن شامل چهار سطحِ پردازش لغوی، پردازش ساختواژی، پردازش نحوی و پردازش معنایی می‌شود که عملیات ریشه‌یابی در سطح پردازش لغوی صورت می‌گیرد. در زبان‌های طبیعی از بین کل واژه‌ها تعداد معدودی واژه یافت می‌شود که واژه‌های ریشه در نظر گرفته می‌شوند و بقیه‌ی واژه‌ها از این ریشه‌ها مشتق می‌شوند. به دیگر سخن هر واژه دارای ریشه‌ای است که معنا و مفهوم اصلی آن واژه را در بر می‌گیرد. هدف از ریشه‌یابی، زدودن الحاقات و یافتن جوهره اصلی واژه است. ریشه‌یابی به فرایندی اطلاق می‌شود که در آن با حذف پیشوندها و پسوندها نهایتاً ریشه‌ی یک واژه مشخص می‌شود. به عبارت دیگر بعد از انجام عملیات ریشه‌یابی، واژه‌ی حاصل قابل تجزیه نخواهد بود. هنگامی که این عملیات توسط رایانه انجام شود به آن ریشه‌یابی خودکار گفته می‌شود.

به‌طور کلی عملیات ریشه‌یابی کلمات در دو دسته‌ی ساختاری و غیر ساختاری قرار می‌گیرد. این پژوهش در حوزه‌ی الگوریتم ساختاری قرار دارد. در الگوریتم پیشنهادی از یک روش ترکیبی استفاده می‌شود که ترکیبی از روش حذف پیشوند و پسوند بر اساس حروف پایانی کلمات، الگوریتم بن، روش جدولی و نیز پاره‌ای از قوانین موجود می‌باشد. ابزار بکار رفته در این پژوهش، زبان برنامه نویسی PHP به همراه استفاده از پایگاه داده‌ی MySql است که برای نوشتن برنامه‌ی اصلی بکار رفته است.

AJAX تکنیک دیگری است که برای جلوگیری از refresh شدن صفحه بکار رفته است.

تکنیک دیگر HTML است که مخفف Hyper Text Markup Language می‌باشد

و به معنی زبان نشانه‌گذاری فرا متنی است و برای نشانه‌گذاری عناصر یک صفحه وب به کار می‌رود به طوری که یک مرورگر وب بتواند آن صفحه را به عناصر قابل رویت ترجمه کرده و آن را روی صفحه نمایشگر نمایش دهد. یک زبان نشانه‌گذاری از عناصر علامت گذاری (Tag) ساخته شده است.

از دیگر تکنیک‌ها می‌توان به CSS اشاره کرد که مخفف Cascading Style Sheets جبه معنی شیوه‌نامه‌ی آبشاری است. Styles یا قالب‌ها مشخص می‌کنند که عناصر HTML چگونه نمایش داده شوند.

از دیگر نتایج بدست آمده از این پژوهش می‌توان به ریشه‌یابی اکثر کلمات مرکب در زبان فارسی اشاره کرد؛ نیز این ریشه‌یاب قادر به ریشه‌یابی برخی کلمات عربی دخیل در فارسی بر اساس قوانین پیشنهادی می‌باشد.

منابع

[۱] شمس‌فرد، مهرنوش،۱۳۸۵. «پردازش متون فارسی: دستاوردهای گذشته، چالشهای پیش رو»، *دومین کارگاه پژوهشی زبان فارسی و رایانه*: ۱۷۲- ۱۸۹.

[۲] تشکری، مسعود، میبدی، محمدرضا، ۱۳۸۰. «طراحی یک ریشه‌یاب خودکار برای واژگان فارسی»، *مجموعه مقالات هفتمین کنفرانس سالانه انجمن کامپیوتر ایران*.

[۳] عباسی، محسن، منصفی، رضا، استیری، احمد،۱۳۹۱. «طراحی یک سیستم توصیه‌گر ترکیبی معنایی با استفاده از تکنیک‌های پردازش زبان طبیعی فارسی»، *نخستین کنفرانس بین المللی پردازش خط و زبان فارسی*، دانشگاه سمنان، دانشکده برق و کامپیوتر.

[۴] زاهدی، محمد صادق، بزرگی، ارسطو، فاتحی، کاوان، ۱۳۹۱. «بررسی ریشه‌یاب‌های واژگان زبان فارسی و تأثیر آنها در کارایی سیستم‌های بازیابی اطلاعات متنی»، *نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی*، دانشگاه سمنان، دانشکده برق و کامپیوتر.

[۵] یوسفیان، احمد، طباطبایی، بی‌بی صدیقه، ۱۳۹۱. «پیاده‌سازی یک غلط‌یاب املایی فارسی تحت وب»، *نخستین کنفرانس بین المللی پردازش خط و زبان فارسی*، دانشگاه سمنان، دانشکده برق و کامپیوتر.

[۶] مشکی، محسن، آنالویی، مرتضی، ۱۳۸۶. «خلاصه سازی چند سندی متون فارسی با

استفاده از یک روش مبتنی بر خوشه بندی»، *اولین کنفرانس ملی مهندسی نرم افزار ایران*.

[۷] کاشفی، امید، نصری، میترا، کنعانی، کامیار، ۱۳۸۹. «خطایابی املایی خودکار در زبان فارسی»، دبیرخانه *شورای عالی اطلاع رسانی*.

[۸] محمدی نصیری، مجتبی، شیخ اسماعیلی، کیومرث، ابوالحسنی، حسن، ۱۳۸۴. «یک ریشه یاب آماری برای زبان فارسی»، *یازدهمین کنفرانس بین المللی کامپیوتر انجمن کامپیوتر ایران، کنفرانس CSICC*.

[۹] کامیار، محسن، ۱۳۸۷. «ریشه‌یابی کلمات فارسی»، *سمینار هفتگی دانشگاه فردوسی مشهد*.

[10] Frakes,W. &R, Baeza-Yates,1992. "Information Retrieval: Data Structures& Algorithms,Stemming Algorithms".

[۱۱] احسان، نوا، فیلی، هشام،۱۳۹۰. «بررسی تأثیرات ریشه‌یابی در بازیابی اطلاعات در زبان فارسی»، *دوفصلنامه تخصصی پردازش علائم و داده‌ها*، سال هشتم، شماره ۱: ۱۷-۲۴.

یادگیری ساختار دستوری زبان انگلیسی در مدارس با استفاده از امکانات چند رسانه‌ای

صدیقه سادات مقداری* و **فاطمه علوی شهری****

چکیده

کاربرد صحیح فناوری‌های جدید مثل رایانه و تجهیزات وابسته به آن به خصوص چند رسانه‌ای‌ها در آموزش و یادگیری اهمیت بسزایی دارد و به بهبود فرایند یاددهی و یادگیری کمک می‌کند.هدف از انجام پژوهش حاضر تعیین تاثیر آموزش چند رسانه‌ای در مقایسه با روش سنتی بر پیشرفت تحصیلی و ایجاد انگیزه و عملکرد مناسب در زمینه مهارت‌های زبان انگلیسی دانش آموزان پایه دهم می‌باشد. جامعه آماری پژوهش حاضر عبارت است از کلیه دانش آموزان پایه دهم دبیرستان محمودیه گناباد که در سال تحصیلی ۹۶-۹۵ مشغول تحصیل بودند و تعداد آنها ۱۰۰ نفر بودند. حجم نمونه برای هر گروه ۲۲ نفر (گروه آزمایش و گروه کنترل) که به روش نمونه گیری تصادفی از بین کلاس‌های پایه دهم دبیرستان انتخاب شدند. در پژوهش حاضر روش تحقیق نیمه آزمایشی می‌باشد. به منظور گرد آوری داده‌ها از آزمون پیشرفت تحصیلی معلم ساخته به صورت پیش آزمون و پس آزمون و پرسشنامه استفاده شد. داده‌های تحقیق با استفاده از شاخص‌های آمارتوصیفی شامل میانگین و انحراف معیار تجزیه و دسته بندی شد. فرضیه تحقیق با مقایسه نمرات پس آزمون و پیش آزمون گروه‌های کنترل و آزمایش با استفاده از آزمون تی و با کمک نرم افزار تحلیل گردید. تفاوت‌های چشمگیری بین دو گروه از نظر نگرش و کنش فراگیران وجود داشت. این پژوهش نشان داد که استفاده از چند رسانه‌ای آموزشی محقق ساخته بر یادگیری جنبه‌های مختلف زبان انگلیسی از جمله (واژگان و اصطلاحات –درک مطلب و ساختار جملات) دانش آموزان پایه دهم دبیرستان تاثیر مثبت داشته است.

واژه‌های کلیدی: چند رسانه‌ای، یادگیری، چند رسانه‌ای محقق ساخته، آموزش سنتی، پیشرفت تحصیلی، مواد آموزشی با کمک رایانه.

---

* استادیار زبانشناسی همگانی، دانشگاه پیام نور، s_meghdari@pnu.ac.ir

** کارشناسی ارشد زبانشناسی همگانی و دبیر آموزش و پرورش، parisaalavi77@gmail.com

۱. مقدمه

حرکت آموزش و فراگیری از مواد آموزشی، کاغذ محور به سویه ی مواد آموزشی رایانه محور شتابی روزافزون می‌گیرد به طوری که یادگیرندگان به استفاده از فناوری در یادگیری و ابزاری مانند رایانامه و اینترنت سوق یافته اند. مواد آموزشی رایانه محور، نمودهای یادگیری زبان به کمک رایانه محسوب می‌شود. یادگیری زبان به کمک رایانه آموزش زبان حضوری نزد معلم را کامل می‌کند تا اینکه جایگزین معلم شود. از ابزارهای رایانه محور می‌توان در ارزشیابی پویا هم استفاده کرد که به یادگیرندگان کمک می‌کند تا در فرایند یادگیری کارکرد بهتری داشته باشند و سطح یادگیریشان را ارتقا دهند. در کشورهای پیشرفته هم، بر خلاف تصور غالب، هنوز استفاده از ابزار آموزشی رایانه محور گسترش چندانی نیافته است که با بررسی‌های بیشتر می‌توان دلایل آن را ریشه یابی کرد. . آنزوزث (۲۰۰۸:۲۲) [۱] برای این منظور از دو راهکار همگونی (assimilation) و همزیستی (accomodation) استفاده می‌کند. به نظر وی معلمان باید مفاهیم علم آموزی جدید را در آموزش‌هایشان همگون سازند به این معنی که باید در ارتباط با علم آموزی معمول آنها را مفهوم سازی کرده و در امور آموزشی آنها را اجراکنند. همچنین علم آموزی جدید باید با آموزش علم همزیست شود. یعنی مفاهیم و تجربیات علم آموزی جدید باید به بازسازی اساسی این فکر منجر شود که طبیعت و نحوه ی علم آموزی باید کاملا عوض شود. یعنی منظور از همزیستی کاربرد مکمل آن و نه نابود کردن علم آموزی معمول می‌باشد. لذا همگونی و همزیستی به معنی تکمیل کردن آموزش با استفاده از متون الکترونیک و سایر نظام‌های نشانه شناختی در کنار سیستم زبانی واژه – محور می‌باشد. لذا علم آموزی جدید باید با توجه به طبیعت خودش و در ادامه ی علم آموزی تثبیت شده مورد بحث و بررسی واقع شود و مفهوم سازی مجددی از مقوله ی علم آموزی متناسب با عصر چند رسانه-ای دیجیتال صورت گیرد. لذا کرس(۱۵۵:۲۰۰۰) [۲] اذعان می‌دارد که " نظریه ی نشانه شناسی که توجهی به این تغییرات اساسی نکند در عصر حاضر ناقص و ناکارآمد خواهد بود." چراکه یکی از اصول اساسی برای موثر بودن آموزش این است که روش‌های آموزشی،عناصر و رسانه‌ای ارایه دهنده ی آن باید به فراگیر کمک کند تا فرایند یادگیری و درک اطلاعات و مهارت‌ها بطور موثری انجام پذیرد.

انسان موجودی اجتماعی است و برای رفع نیازهای خود باید با دیگران ارتباط برقرار کند

و بدون ارتباط هیچ فعالیت اجتماعی بین افراد شکل نمی گیرد. امروزه با ورود فناوری‌های نوین آموزشی و تحول در آموزش زبان انگلیسی در قالب ارایه مجموعه کتاب‌های English for schools از سال تحصیلی ۱۳۹۲-۱۳۹۱ آغاز گشت. مجموعه مذکور دوره‌ای شش جلدی شامل دو زیر مجموعه ۳ جلدی با نام های prospect برای متوسطه اول و vision برای متوسطه دوم می‌باشد. ویژگی رویکرد ارتباطی فعال و خود باورانه و چند رسانه‌ای متکی بر اصول کلی زیر است.

۱- توجه هم زمان به هر چهار مهارت زبانی

۲- استفاده از فعالیتهای آموزشی سمعی و بصری و چند رسانه‌ای متنوع در فرآیند یادگیری زبان

۳- ارتقای روحیه فراگیری زبان در محیط مشارکتی و از طریق همکاری و همیاری در کلاس

۴- ارایه بازخوردهای اصلاحی به خطاهای فراگیران. با کاربرد کامپیوتر در آموزش، برنامه درسی با شرایط و ویژگی‌های فراگیر تنظیم می‌شود و مشکل تفاوت‌های فردی که از دیر باز در تعلیم و تربیت مطرح بوده کاهش می‌یابد. بدین معنی که اگر در آموزشهای سنتی معلمان فرصت کافی برای شناخت دانش آموزان و کارکردن انفرادی با آنها را ندارند، کامپیوتر می‌تواند فرصت‌ها و تجارب یادگیری متفاوت و متنوعی در اختیار فراگیران قرار دهد. با صرف زمان بیشتری در مورد فراگیرانی که مشکلات یادگیری دارند آنها را به سطح مطلوب رسانده و بدین طریق مشکلات فردی در آموزش را برطرف نماید. همچنین کامپیوتر محدودیت‌های زمانی و مکانی را در هم شکسته و آموزش را در هر زمان و مکانی امکان پذیر سازد (سعادتمند و همکاران، ۱۳۹۱) [۱]. یک برنامه چند رسانه‌ای محقق ساخته خوب طراحی شده، می‌تواند به یاد گیرنده درباره موضوعات درسی جدیدآگاهی دهد، تمرینات و تجارب کافی در اختیار فراگیر قرار دهد، مرور مهارتهای لازم که برای موفقیت دانش آموزان لازم است را امکان پذیر می‌کند، بازخورد مناسب ارایه دهد و پیشرفت فراگیر را بطور مداوم ارزشیابی کند. با توجه به این که در بازار ایران چندین سال است که تولید چند رسانه‌ای‌ها در دستور کار وزارت آموزش و پرورش و شرکت‌های خصوصی قرار گرفته است، از طرفی علاوه بر کیفیت، هزینه نیز مطرح می‌باشد که وجود استانداردها و تبعیت از آن‌ها سبب می‌شود به جای تولید هزاران قطعه محتوای

الکترونیکی مشابه به تعداد بسیار محدودی از قطعات تولید شده یا در حال ساخت اکتفا و سپس از ترکیب این قطعات محتوایی، دروس یا دوره‌های مختلف و متنوعی ایجاد شود (شاه جعفری، ۱۳۹۲) [۲] .

هنگام استفاده از کامپیوتر و یادگیری از طریق نرم افزار آموزشی، توجه فراگیران به جای تخته سیاه قدیمی کلاس و گچ، معلم و دیگر همکلاسی‌ها جلب صفحه کامپیوتر می‌شود و همین امر سبب تمرکز، تفکر و در نهایت عکس العمل بهتر و سریع تر آنان می‌شود (بهرنگی و اسدی ۱۳۹۱) [۳] .

تدریس یک فرایند است و فعالیتی است که در داخل یک الگو صورت می‌گیرد. الگوی تدریس چند رسانه‌ای در مقایسه با الگوی سنتی چهار چوب ویژه‌ای دارد که عناصر مهم تدریس در درون آن قابل مطالعه است. انتخاب یک الگوی تدریس بستگی به نوع آگاهی معلم از فلسفه و نگرش‌های تعلیم و تربیت خواهدداشت .

یکی از ویژگی منحصر به فرد کامپیوتر در آموزش که ماشین آموزش نام گرفته دقت و سرعت عملی است که در آموزش و یادگیری دارند و از نقاط ضعف نیروی انسانی که در آموزش سنتی به کار رفته مانند خستگی ,فراموشی و سایر عللی که باعث افت بازده آموزشی می‌شود مبرا هستند (سعادتمند و همکاران ۱۳۸۱) [۱] .

از محاسن استفاده از چند رسانه‌ای‌ها در تدریس می‌توان به موارد زیر اشاره کرد.

۱- پرسش‌ها به صورت زنجیرهای به یکدیگر وابسته اند.
۲- مفاهیم با یک سیر منطقی در آن تنظیم شده است.
۳- ماشین بر خلاف انسان دچار عوارضی مانند بی حوصلگی، عصبانیت و ناراحتی نمی شود.

آموزش به کمک رایانه با قابلیت چند رسانه‌ای می‌تواند حواس گوناگون را همزمان در فرآیند تجربه چند حسی به کار گیرد و برای افراد با ویژگی‌های متفاوت محیط مطلوب یادگیری ایجاد نماید (عالمی ۱۳۸۹) [۴].

۷۵ درصد یادگیری از طریق وسایل دیداری و تصویری به وسیله حس بینایی انجام می‌شود. در صورتی که تنها ۱۳ درصد یادگیری از طریق حس شنوایی وسایل صوتی انجام می‌گیرد و دیگر حواس به ترتیب بساوایی ۶ درصد، بویایی و چشایی هر کدام ۳ درصد در

حافظه یادگیری تاثیر دارد و فرایند یادگیری کامل را متاثر و اثر بخش می‌نمایند (ابراهیمی، ۱۳۹۲) [۵] .

۲. پیشینه تحقیق

کاربرد فناوری در آموزش موضوع تازه‌ای نیست، اما کاربرد فناوری در یادگیری زبان برای یادگیرندگان، معلمان و دانشمندان در مراحل اولیه است. مفاهیم پیچیده‌ای که امروزه دانشجویان و حتی دانش‌آموزان با آن روبرو می‌باشند ضرورت توسعه ی مفهوم " علم آموزی تک رسانه‌ای " را به یک مفهوم تجمیعی از علم آموزی نشان می‌دهد که شامل علم آموزی بصری، کلامی و سایر موارد می‌شود. این مفهوم " علم آموزی چندگانه" یا علم آموزی الکترونیک " نامیده می‌شود.همچنین اصطلاحات " علم آموزی دیجیتال" ، " علم آموزی مجازی " ، " علم آموزی جدید" و " علم آموزی ابررسانه‌ای " نیز نامیده شده است (ر.ک. انزورث، ۲۰۰۶)] ۳[. لذا بسیار ی از صاحبنظران و نظزیه پردازان به ارائه ی نظریاتی تحت عنوان زبان تصویر پرداخته اند. از جمله شناخته شده ترین آنها کتاب " دستور زبان طرح‌های بصری" اثر کرس و ون لیون (۲۰۰۶) می‌باشد. کوپ و کالانتیز (۲۰۰۰) [۴] نیز بر اهمیت تصاویر بعنوان منبع معنی سازی در متون چند رسانه‌ای اخیر تاکید کرده اند. و اینکه در محیط‌های مجازی معنی سازی زبانی و بصری- نموداری با هم ترکیب می‌شوند. لذا هرچند در محیط‌های رسانه‌ای کاغذی و الکترونیکی اصول اساسی خواندن و نوشتن تغییر نکرده است اما پردازش آنها از پردازشی شناختی و سریالی متون خطی به پردازشی موازی متون تصویری تغییر کرده است.

برنامه‌های آموزشی با کمک رایانه، به داربستی می‌ماند که به زبان آموزان بازخوردهای لازم را ارائه می‌دهد و به زبان آموز فرصت می‌دهد که با استفاده از خودآگاهی و مشاهده خود در حین فرایند یادگیری به سوی مستقل شدن در یادگیری پیش برود. تازگی این پژوهش پرداختن به عوامل عاطفی در یادگیری زبان با کمک رایانه بود . عواملی هم چون انگیزه دادن، افزایش خودکارایی، جالب بودن و چالش انگیز بودن که این عوامل خود نقش سازنده و حمایتی معلمان را آشکارتر می‌کند. در واقع ما از نظر فرهنگی به ابزار رایانه‌ای و فناوری به عنوان عناصری سرد و بی روح نگاه می‌کنیم، اما گنجاندن برنامه‌ها و ابزار هیجان انگیز و چالشی

می‌تواند به این ابزار روح تازه‌ای دهد.

پژوهشی که توسط عباسی و همکاران (۱۳۸۸) [۶]با عنوان، مشکلات یاد دهی و یاد گیری درس زبان انگلیسی و استفاده از روش‌های سنتی آموزش انجام گرفت به این نتیجه رسیدند که این عامل ممکن است ناشی از :

۱- ضعف محتوای کتب درسی

۲- تناسب نداشتن متون کتب درسی با توان دانش آموزان

۳- استفاده نکردن از روش‌های تدریس متنوع و فعال

۴- کمبود ساعات اختصاص داده شده به درس زبان

۵- استفاده نکردن از چند رسانه‌ای‌ها و وسایل سمعی و بصری

۶- تسلط نداشتن معلم به مهارت‌های دانش زبان انگلیسی و ناکارآمدی در استفاده از نرم افزار مولتی مدیا در کلاس درس

۳. اهداف پژوهش

این پژوهش مبتنی بر اهداف و متغیرهای مورد مطالعه از نوع پژوهش‌های نیمه تجربی با طرح پیش آزمون و پس آزمون بین دو گروه کنترل و آزمایش است که در آن اثرات یک متغیر مستقل (آموزش چند رسانه ای) بر متغیر وابسته (یاد گیری زبان انگلیسی) مورد بررسی قرار خواهد گرفت. در پژوهش حاضر بررسی تاثیر آموزش چند رسانه‌ای در مقایسه با روش سنتی بر یادگیری زبان انگلیسی از جنبه‌های درک مطلب و لغات، اصطلاحات و ساختار جملات مورد بحث قرار می‌گیرد.

۴. روش تحقیق

در این تحقیق پس از بررسی‌های لازم به علت هم سطح بودن دانش آموزان، شرایط و امکانات و هم جنس بودن طرح در بزرگترین دبیرستان دخترانه که دارای ۳ کلاس پایه دهم با تعداد ۱۰۰ نفر دانش آموز بود اجرا شد. پرسشنامه کتاب جدید التالیف توسط محقق تنظیم و در بین فراگیران پایه دهم توزیع گردید. بعد از تکمیل و ارزیابی دو گروه آزمودنی هر یک به تعداد ۲۲ نفردر نظر گرفته شدند. در مرحله بعد جهت سنجش پیش دانسته‌های دانش آموزان

آزمون پیشرفت تحصیلی از کتاب انگلیسی پایه دهم متوسطه که از نوع محقق ساخته بود و به تایید همکاران زبان انگلیسی هم که در این حوزه فعالیت داشته رسیده بود به صورت پیش آزمون برگزار گردید و مشخص شد که هر دو گروه در ابتدا از لحاظ توانایی زبانی هم سطح می‌باشند. یک گروه به عنوان گروه کنترل با روش سنتی آموزش داده شد و یک گروه به عنوان گروه آزمایش تحت تاثیر متغیر مستقل یعنی تاثیر آموزش به کمک تجهیزات چند رسانه‌ای بر یاد گیری زبان انگلیسی قرار گرفت. در اولین جلسات به علت عدم تسلط بعضی از فراگیران به کار با این تجهیزات و اهداف آنها کار به کندی پیش می‌رفت و باید آموزش فناوری ارتباطی صورت می‌گرفت. محتویات کتاب جدید التالیف پایه دهم بر اساس ارتقای روحیه ی فراگیری و ایجاد انگیزه در محیط مشارکتی و مبتنی بر تجهیزات چند رسانه‌ای از طریق همیاری و همکاری در کلاس درس است. متن‌ها و لغات که در دو گروه تدریس شدند از کتاب درسی پایه دهم انتخاب شدند.

فعالیت‌های زیر در گروهی انجام می‌گرفت که به روش سنتی آموزش می‌دیدند:

۱- روخوانی متن توسط معلم

۲- تکرارآن توسط فراگیران جهت تسلط در امر تلفظ و یادگیری

۳- ترجمه متن به زبان فارسی یا زبان مادری

۴- حفظ عبارات و ساختارها

۵- کلاس کامل معلم محور است و به روش سخنرانی اداره می‌شود و دانش آموزان شرکت فعالی ندارند.

فعالیت‌های زیر در گروهی انجام می‌گرفت که به روش چند رسانه‌ای آموزش می‌دیدند:

۱- برای تدریس لغات جدید فقط استفاده از زبان انگلیسی بود و به صورت نمایش و استفاده از اشیای واقعی در کلاس درس بود.

۲- استفاده از اشارات (gestures) یا نمایش مراحل (acting out)

۳- استفاده از فلش کارت (flash cards) به صورت موارد موجود در بازار یا معلم ساخته.

۴- نمایش تعاریف، مترادف و مخالف واژگان

۵- کاربرد واژگان در جمله از نظر نقش دستوری و نمایش مثال‌های عینی برای کلمات

۶- استفاده از فرهنگ لغت برخط روی سیستم

۷- تهیه پیکره در کلاس درس زبان که مستلزم زمان و برنامه ریزی دقیق می‌باشد و تمامی فعالیت‌های آن مستلزم دسترسی به کامپیوتر است و با استفاده از فهرست‌های تکرار ساده و خروجی فهرست الفبایی انجام می‌شود. در این کلاس از فهرست الفبایی کلمات هدف برای فهم بهتر کلمات و کشف ارتباط معنایی آنها کمک گرفته شد. این امر باعث شد یاد گیری زبان راحت تر و درک الگوها برای زبان آموزان آسانتر شود.

برای تدریس قسمت خواندن و درک مطلب ۳ مرحله اساسی در این کلاس‌ها را انجام شد:

۱- مرحله آمادگی (pre-reading) که در این مرحله فراگیران را با پیام اصلی (theme) درس یعنی همان هدف کلی درس آشنا می‌کنیم با نمایش چند اسلاید مرتب با متن درس ذهن فراگیران را برای مطلب مورد آموزش آماده سازی می‌کنیم.

۲- برای شرکت فعال و پویای فراگیران و یادگیری از راه تعاملی و ارتباط متقابل از روش بارش فکری (brain storming) و جمع آوری نظرها و افکار فرگیران پیرامون موضوع مورد آموزش.

۳- مرحله خواندن (while-Reading) در حین اجرای آزمون تعدادی سوال روی تخته می‌نویسیم تا فراگیران حین گوش دادن به فایل صوتی درس دنبال پاسخ سوالات نوشته شده باشند.

۴- مرحله پساخواندن (post-Reading) از فراگیران خواسته می‌شود که ایده اصلی متن یاد گرفته شده را ذکر کنند. در این مرحله فراگیران باید بتوانند از همدیگر به طور مشارکتی وفعالانه سوال و جواب کنند.

۵- در آخرین مرحله بازیهای رایانه‌ای به صورت جورچین طراحی شده و دانش آموزان کلمات را جایگزین می‌کنند. انجام فعالیت صحبت کردن در کلاس مرتبط با مطلب آموزش داده شده به صورت تک گویی یا مکالمه‌ای و تهیه پاورپوینت در مورد ساختارهای گرامری، لغات و اصطلاحات توسط فراگیران.

شیوه نمره گذاری در روش سنتی

۱- روخوانی صحیح

۲- ترجمه عبارات از انگلیسی به فارسی و بر عکس

۳- حفظ فرمول‌ها و ساختارهای جمله سازی و زمانها

در روش سنتی تاکید روی خواندن و نوشتن است و توانایی ارتباط برقرار کردن به طور شفاهی با زبان دوم کم است.

۵. شیوه نمره گذاری در روش چند رسانه‌ای

۱- خواندن به صورت انفرادی فقط با چشم برای تسلط در تلفظ و درک و فهم متن از طریق ساکت خوانی (silent Reading) و توجه به با هم آیی کلمات و عبارات در هر پاراگراف.

۲- فراگیران باید متن را کامل بفهمند و گزینه صحیح را انتخاب کنند. سوالات به طور زمان دار مطرح شده و بعد از زمان خواسته شده سیستم قطع می‌شود و نمره اعلام می‌گردد بصورت سوال چند گزینه‌ای(Multiple choice question)

۳- متن مرتبط با متون کتاب درسی داده شده اما در سطح فراتر که در زمان مشخص خوانده شده و به سوالات آن به صورت درست یا غلط(True or False) پاسخ داده شود.

۴- فراگیران برای فهم کامل کلمات هر یک را به تعریف صحیح آن متصل می‌کنند (Matching exercise).

آموزش ۴ مهارت زبانی موضوع مهمی است در آموزش زبان دوم

	شفاهی	کتبی
دریافتی	گوش کردن	خواندن
تولیدی	صحبت کردن	نوشتن

در روش سنتی مهارت‌های تولیدی کم رنگ و ضعیف است و در روش چند رسانه‌ای هر ۴ مهارت به طور همزمان کار می‌شود و فراگیران از الگوهای صدا و کلمات و شروع کننده‌های

دستور زبان برای تولید ساختار زبان و نوشتن آن استفاده می‌کنند. طبق نظریه تعامل گرایان اجتماعی بیشتر رشد زبان شناختی بچه‌ها از الگوگیری و تعامل با یکدیگر و تصحیح‌های آموزشی آنها ناشی می‌شود (ناعمی ۱۳۸۵) [۷] .

به این ترتیب آموزش به کمک چندرسانه‌ای‌ها و شیوه سنتی به مدت ۱۰ هفته ادامه داشت و در پایان آزمون پیشرفت تحصیلی از درس‌های داده شده به صورت پس آزمون گرفته شد.

۶. روایی و پایایی آزمون

برای روایی این آزمون از همکاران زبان که مشغول تدریس در این پایه بودند کمک گرفته شد تا سوالات طراحی شده با توجه به دو بعد هدف و محتوا به منظور ارزشیابی میزان یادگیری دانش آموزان از مطالب کتاب باشد. برای پایایی آزمون مذکور از تعداد زیاد سوالات متجانس در سطح متوسط استفاده شد. آزمون به صورت ۴ گزینه‌ای تا برای نمره گذاری آسان و عینی مناسب تر باشد و حدس زدن تاثیر کمی روی آن داشته باشد. بر اساس نظریه‌های یادگیری گستره‌ای از تفاوت‌ها از قبیل سن، استعداد، اولویت سبک یادگیری و استفاده از استراتژی و انگیزه که می‌تواند سرعت یادگیری زبان دوم را تحت تاثیر قرار دهد.

۷. یافته ها:

در این بخش اطلاعات و داده‌های جمع آوری شده در دو گروه کنترل و آزمایش در دو مرحله پیش آزمون و پس آزمون با استفاده از آزمون t وتحلیل کواریانس مورد تجزیه و تحلیل قرار گرفت.

فرضیه ۱: استفاده از نرم افزار چند رسانه‌ای محقق ساخته بر یادگیری درس زبان انگلیسی پایه دهم دبیرستان تاثیر مثبت دارد.

جدول ۱: مقایسه نمرات پیش آزمون مهارت‌های یادگیری زبان انگلیسی در گروه‌های کنترل و آزمایش

متغیر	گروه	تعداد	میانگین	انحراف استاندار	آزمون T	درجه آزادی	سطح معنی داری
واژگان	کنترل	۲۲	۷/۶۳۶۴	۲/۷۶۱۰۵	۰/۴۶۰	۴۲	۰/۶۴۸
	آزمایش	۲۲	۷/۹۵۴۵	۱/۷۰۳۷۰			
درک مطلب	کنترل	۲۲	۸/۵۴۵۵	۱/۷۶۵۴۷	۰/۹۷۲	۴۲	۰/۳۳۷
	آزمایش	۲۲	۸/۰۰۰۰	۱/۹۵۱۸۰			

بر اساس اطلاعات جدول در هر دو مهارت سطح معنی داری از ۰/۰۵ بالاتر است. نتیجه می‌شود تفاوت نمرات پیش آزمون این دو مهارت یادگیری زبان بر اساس گروه‌های مورد مطالعه معنی دار نیست یعنی دو گروه قبل از اجرای آزمایش تقریبا یکسان بوده اند.

فرضیه ۲: آموزش به کمک چند رسانه‌ای در یادگیری مهارت درک مطلب زبان انگلیسی موثر است.

جدول توزیع پراکندگی نمرات مهارت درک مطلب در دو گروه آزمایش و کنترل

گروه	تعداد	میانگین	انحراف استاندار
کنترل	۲۲	۸/۴۰۹۱	۱/۷۰۸۷۸
آزمایش	۲۲	۹/۱۸۱۸	۰/۹۵۷۹۹

جدول نتایج تحلیل کواریانس نمرات مهارت درک مطلب بین گروه کنترل و آزمایش با کنترل نمرات پیش آزمون

منبع	مجموع مجذورات	درجه آزادی	میانگین مجذورات	آزمون f	سطح معنی داری
مهارت درک مطلب (پیش آزمون)	۴۳/۰۰۶	۱	۴۳/۰۰۶	۴۶/۹۱۳	۰/۰۰۰
گروه ها	۱۲/۳۰۱	۱	۱۲/۳۰۱	۱۳/۴۱۹	۰/۰۰۱

تفاوت مشاهده شده بین دو گروه از نظر مهارت درک مطلب معنی دار است و در کل میانگین نمرات در گروه آزمایش بالاتر است. پس نتیجه می‌شود که آموزش به کمک رایانه در یادگیری مهارت درک مطلب زبان انگلیسی موثر است.

فرضیه ۳: آموزش به کمک چند رسانه‌ای‌ها در یادگیری لغات و اصلاحات زبان انگلیسی موثر است.

توزیع پراکندگی نمرات مهارت لغات و اصطلاحات در دو گروه کنترل و آزمایش

گروه	تعداد	میانگین	انحراف استاندارد
کنترل	۲۲	۷/۹۷۷۳	۱/۹۵۴۷۱
آزمایش	۲۲	۹/۴۰۹۱	۰/۷۱۷۷۴

نتایج تحلیل کواریانس نمرات مهارت لغات و اصطلاحات بین گروه کنترل و آزمایش با کنترل نمرات پیش آزمون

منبع	مجموع مجذورات	درجه آزادی	میانگین مجذورات	آزمون f	سطح معنی داری
مهارت لغات (پیش آزمون)	۴۲/۷۶۶	۱	۴۲/۷۶۶	۳۶/۳۰۹	۰/۰۰۰
گروه ها	۱۸/۲۶۶	۱	۱۸/۲۶۶	۱۵/۵۰۸	۰/۰۰۰

تفاوت مشاهده شده بین دو گروه از نظر مهارت لغات و اصطلاحات معنی دار است و در کل میانگین نمرات در گروه آزمایش بالاتر است. نتیجه می‌شود آموزش براساس رایانه در یادگیری و مهارت لغات و اصطلاحات زبان انگلیسی موثر بوده است.

استفاده از واژه پرداز برای نوشتن و استفاده از ساختارهای دستوری نگرش یادگیرندگان زبان دوم را بهبود می‌بخشد که می‌توانند تا حدودی به این حقیقت برسند که ترس دانش آموزان برای اشتباه کردن و سپس دوباره نوشتن کل متن کمتر می‌شود و آن‌ها را از دوباره نویسی کل متن برای دو یا چند بار آسوده می‌کند.

بحث و نتیجه گیری

کشور ما، خوشبختانه پتانسیل استفاده از ابزارهای آموزشی با کمک رایانه را دارد و ابزارهای فناوری در زندگی فردی یادگیرندگان عجین شده اند و از این رو می‌توان از این ابزارها در جهت مثبت و برای تقویت زبان انگلیسی استفاده کرد و با بکارگیری آن‌ها ، یادگیرندگان را در معرض هر چه بیشتر زبان قرار داد.

همان طور که بحث شد آموزش به کمک رایانه می‌تواند به یادگیرنده در مورد موضوعات درسی آگاهی دهد، تمرینات و تجارب کافی در اختیار دانش آموزان قرار دهد و فراگیر را درگیر فرایند یاد گیری نموده و بالطبع یادگیری فعال آنان را موجب شود. کولی، کرادور و وانگل خاطر نشان کردند که فناوری می‌تواند برای از میان برداشتن فرصت‌ها ی نا برابر دانش آموزان با پیش زمینه‌های تحصیلی گوناگون به آنان امکان بدهد تا از گنجینه اطلاعات موجود در شبکه اینترنت بطور مساوی به مساوی بهره مند شوند (صیف و بیرانوند ۱۳۸۸) [۸] .

در این پژوهش فرضیه دوم یعنی تاثیر آموزش به کمک چند رسانهای بر یادگیری درک مطلب تایید شد. با توجه به نتایج تفاوت میانگین دو گروه معنی دار بود. این یافته‌ها با یافته‌های پژوهش‌های دیگر نیز همخوان است. نتایج یافته‌های جعفری کوخالو و حمیدی نشان داد که کارکردهای شناختی بالای دانش آموزان (ارزشیابی وتجزیه و ترکیب) درهم آموزی بیشتر از سنتی است. در تبیین این یافته می‌توان گفت، با توجه به اینکه در آموزش چند رسانه‌ای فراگیران به صورت یک گیرنده منفعل عمل نمی کنند، بلکه مشتاقانه با راهنمایی معلم به نقاط ضعف و قوت خود پی می‌برند و همین باعث تعیین مسیر یادگیری می‌گردد. لذا یادگیری مهارت درک مطلب به آسانی صورت می‌گیرد.

نتایج فرضیه سوم نشا ن داد که میانگین نمرات فراگیرانی که در یادگیری لغات و اصطلاحات زبان انگلیسی در گروهی که به کمک نرم افزارهای چند رسانه‌ای آموزش دیده بودند بالاتر از میانگین نمرات فراگیرانی است که به طور سنتی آموزش دیده اند. این یافته با نتایج پژوهش‌های اسدی و بهرنگی که به این نتیجه رسیدند که کاربرد نرم افزار چند رسانه‌ای باعث افزایش دایره واژگان در گروه آزمایش در مقایسه با گروه کنترل گردید همسو می‌باشد.

راه کار برای توجیه فراگیری از طریق چند رسانه‌ای نسبت به روش سنتی:

۱- به کار گیری هم زمان چند حس بـه طـور آگاهانـه در جریـان یـاد گیـری بـا وجـود مولفه‌های یک برنامه چند رسا نه‌ای مانند صدا، تصویر، حرکت و رنگ.

۲- جدید بودن، تازگی و جنبه جذابیت منحصر به فرد شیوه ارایه مطالب آموزشی بـا بـه کارگیری نرم افزارهای آموزشی که علاقه و توجه فراگیران را جلب کـرده و بـالطبع یـاد گیـری فعال آنها را موجب شده است.

۳- استفاده مجدد از برنامه آموزشی با توجه به قابلیت تکرارپذیری آن.

۴- به یادگیرندگان توصیه می‌شود تا با ابزار رایانه محور بیشتر آشنا شوند و سعی کنند با یادگیرندگان دیگر و معلم خود از طریق این ابزار همکاری کنند.

محدودیت‌های پژوهش:

۱- با توجه به جدید التالیف بودن کتاب درسـی پایـه دهـم و نبـودن نـرم افـزار و بسـته آموزشی حرفه‌ای مناسب از متون کتاب درسی خود محقق نرم افزار مربوطه را تهیه نمود.

۲- یکی دیگر از محدودیت‌های این پژوهش عدم کنترل برخـی از متغیرهـای مـزاحم در طول آموزش بود.

۳- عدم تمایل برخی از دانش آموزان به انجام تکالیف با نرم افزارهای چند رسـانه‌ای بـود که با تشویق محقق نهایتا تکالیف را انجام دادند.

۴- عدم آشنایی معلمان در استفاده از رایانه در کلاس‌های درس.

۵- طفره رفتن از شیوه‌های جدید آموزشی و تمایل به روش‌های سنتی.

۶- نبود وسایل چندرسانه‌ای در کلاس‌های درس.

پیشنهادات:

۱- بهتر است همزمان با چاپ کتاب جدید نـرم افزارهـای مناسـب و مفیـدی در اختیـار دبیران آن رشته قرار بگیرد.

۲- برگزاری دوره‌های ضمن خدمت جهت آموزش مدرسان در کلاسهای درس.

۴- ظتوجه به تفاوت‌های فردی و از روش آموزش چند رسا نه‌ای جهت رشد اسـتعدادهای

خلاق فراگیران استفاده شود.

۴- طراحی بازی‌های رایانه‌ای جهت بهبود عملکرد، ایجاد لذت و شادی، توجه و تمرکز فراگیران

منابع

[1] سعادتمند، محسن، رستگارپور، حسن و فاضلیان، پوراندخت (۱۳۸۱). *مطالعه تاثیر آموزش به کمک کامپیوتر بر یادگیری زبان انگلیسی*، پایان‌نامه کارشناسی‌ارشد رشته تکنولوژی آموزشی، دانشگده روانشناسی و علوم تربیتی تربیت معلم.

[2]شاه جعفری، فاطمه (۱۳۹۰). روش طراحی چند رسانه‌های آموزشی برای ایجاد انگیزه در فراگیران، *نشریه رشد تکنولوژی آموزشی*، شماره۱۶۶.

[3]بهرنگی؛ محمد رضا و اسدی، آرش (۱۳۸۷). همراه سازی نرم افزار مولتی مدیا با الگوی تدریس استقرای نگاره کلمه برای آموزش زبان انگلیسی، *فصلنامه تعلیم و تربیت*. سال بیست و پنجم، شماره ۹۷، ص ص۲۸-۹.

[4] عالمی، محمدحسین (۱۳۸۲). چند رسانه ای‌های آموزشی، *مجله رشد تکنولوژی آموزشی*، دوره پانزدهم، شماره۷، ص ص۱۲-۱۰.

[5] ابراهیمی، زهرا (۱۳۹۲). *مقایسه ی تاثیر رسانه‌های تعاملی بر سرعت و دقت و پایداری یاد گیری*، پایان نامه کارشناسی ارشد، دانشگاه علامه طباطبایی.

[6] عباسی، مهوش ؛ احمدی، غلامرضا و لطفی، احمدرضا (۱۳۸۸). مشکلات یادهی و یادگیری درس زبان انگلیسی دانش آموزان دوره متوسطه اصفهان از دیدگاه دبیران. *پژوهش در برنامه ریزی درسی*. دوره ۲۳، شماره ۲۲.صص ۱۴۱-۱۵۶.

[7]ناعمی، علی‌محمد (۱۳۸۵). *روان‌شناسی آموزش مهارت‌های ارتباط زبان*، مشهد: آستان قدس رضوی.

[8]صیف، محمدحسن و بیرانوند، علی (۱۳۸۸). تاثیر فناوری اطلاعات بر نظام آموزشی مدارس، *مجله اطلاعات علوم و فناوری*.ج.۲۶. صص ۱۸۳- ۱۹۳

[9] Unsworth, L. (2008), *Negotiating New Literacies in English Teaching*. London: Routledge.

[10] Kress, G. (2000), "Design and transformation: New theories of

meaning "in B. Cope and M. Kalantzis (eds), *New Literacies and the English Curriculum*,Melbourne: Macmillan,pp. 153–61

[11] Unsworth, L. (2006). *E-Literature for Children: Enhancing Digital Literacy Learning*. London: Routledge.

[12] Cope, B. and Kalantzis, M. (eds). (2000), *Multiliteracies: Literacy Learning and the Design of Social Futures*. Melbourne: Macmillan.

Translation of Clefting Construction in Persian to English Apertium System

Parya Razmdideh*
Abbas Ali Ahangar**
Seyed Mojtaba Sabbagh-Jafari***

Abstract

Since Persian and English languages differ in several ways including morphology, lexicon, and syntactic structures, this caused machine translation to suffer from numerous problems in translating between these two languages. This paper presents the syntactic challenges of Persian to English Apertium system, a shallow-transfer rule-based machine translation (RBMT) platform, confronted in translating clefted constituents from Persian to English. To this end, 100 sentences containing clefted constituents were gathered in two methods: 1) the contemporary Persian spoken data were collected from radio, television, and daily conversations, 2) the written data were from different Persian grammar books, newspapers, and novels. As the formal words were added to the Apertium dictionaries, it was necessary to pre-process the extracted structures before running them in the Apertium system. To improve the translation performance, new structural transfer rules at *chunker* level were added to the developed system. The Apertium system were evaluated via word error rate (WER) and position-independent error rate (PER). Besides, these sentences were translated using Google translate as a statistical machine translation system (SMT). The evaluation results showed that the Apertium system could generate closer translation to the

* Ph.D. Candidate of Linguistics, University of Sistan and Baluchestan, p.razmdide@gmail.com
** Associate Professor of Linguistics, University of Sistan and Baluchestan, ahangar@english.usb.ac.ir
*** Assistant Professor of Computer Engineering, Vali-e-Asr University of Rafsanjan,mojtaba.sabbagh@vru.ac.ir

reference text comparing that of Google translate due to adding new chunker rules.

Keywords: Rule-based Machine Translation, Apertium, Clefting Construction, Cunker Rules.

1. Introduction

Over the years, machine translation (MT) has been a focus of investigation by linguists, psychologists, philosophers, computer scientists, and engineers. MT is one of the research areas under 'Computational Linguistics'. Various approaches have been devised to automate the translation process. In each of these approaches, the objective has been to restore the meaning of the original text into the translated verse [1]. MT is a difficult task, as many words have various meanings and different possible translations [2]. In some language pairs such as Persian and English, the differences between their phonology, morphology, and syntactic structures give rise to some problems in machine translation. Some of these differences are as follows: Persian is morphologically rich, with many characteristics not shared by other languages, but English belongs to a very simple morphology. Persian makes no use of articles ('a', 'an', 'the') before nouns, but in English depending to the nouns identity (definite or indefinite), a(n) can be used. In addition, there is no distinction between capital and lowercase letters, and symbols as well as abbreviations are rarely used in Persian comparing to English [3]. Persian is a head-final language [4], but for verb-clausal complement order, it is a head-initial one [5] and [6] as cited in [7], and contrary to that, English is a head-initial language [8]. Persian is a pro-drop language unlike English that does not have *pro* [9], it allows a null subject [10]. Also, there are two major differences between the word order in English and Persian. First, English sentences follow the subject-verb-object (SVO) order while Persian sentences follow, in most cases, the subject-object-verb (SOV) order [11]. Second, English has strict word order while Persian allows for free word order. In Persian, the preferred word order is SOV, but all of the other orders

are also correct.

This study considers the syntactic translation problems of clefting construction between Persian and English Apertium shallow-transfer RBMT system. The clefted constituents are translated in the Apertium system (*apertium-pes*[1]*-eng*) by writing new chunker rules. The Apertium system was firstly developed from Persian language to English by authors of the present paper[2].

2. Apertium modules

Apertium[3] is a shallow-transfer and free/open source machine translation system which was published by developers according to GNU GPL (general public license) conditions. Shallow-transfer RBMT systems use relatively simple intermediate representations (IRs), which are based on lexical forms consisting of lemma, part of speech (POS) and morphological inflection information of the words in the input sentence, and apply simple shallow-transfer rules that operate on sequences of lexical forms: this kind of systems do not perform a full parsing [12]. The Apertium shallow-transfer MT platform [13] has recently been used for the development of 47 language pairs[4]. The required linguistic data at Apertium are dictionaries (monolingual and bilingual) and rules (structural transfer and lexical selection). Persian monolingual entries were extracted from the formal and frequent words at Persian side of Mizan English-Persian Parallel Corpus [14]. For the English monolingual dictionary, we used the existing English monolingual dictionary between English and Kazakh language pair [15]. To generate chunker rules, at first, the most common Persian syntactic structures were extracted from different Persian language grammars. These rules are mainly with XML[5]-format (Extensible Markup Language) and hand-written. Apertium platform consists some modules as shown in Figure 1:

1. Persian
2. https://svn.code.sf.net/p/apertium/svn/incubator/apertium-pes-eng
3. http://www.apertium.org
4. http://wiki.apertium.org/wiki/List_of_language_pairs
5. http://www.w3.org/XML/

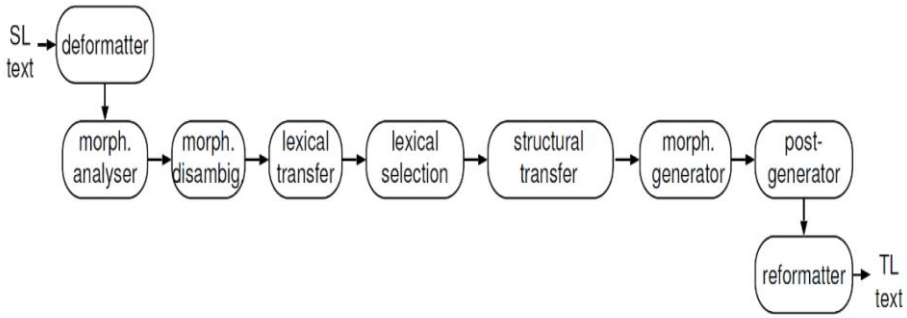


Figure 1. The pipeline architecture of Apertium system [16]

2.1 In **Deformatter**, some formatting tags as HTML (Hyper Text Markup Language) are included in the source text, so they are divided by this module. These tags called as "superblanks" which insert the place between words [13]. For example, a Persian word "اسب" 'horse' with HTML format is like the following example:

1- pes اسب

2.2 Morphological analyser is generated by compiling a morphological dictionary of source language (SL). This tokenizes the text in surface forms (SF) (lexical units as they appear in texts) and delivers, for each SF, one or more lexical forms (LF) consisting of lemma (the base form commonly used in classic dictionary entries), the lexical category (noun, verb, preposition, etc.), and morphological inflection information (number, gender, person, tense, etc.) [17]. For example, the Persian sentence "رضا کتاب را زیر میز گذاشت" 'Reza put the book under the desk' is morphologically analyzed as follows:

2- ^رضا/رضا<np><unk>$ ^کتاب/کتاب<n><sg>$ ^را[1]/را<det><def>$^ زیر/زیر<pr>/$ ^میز/میز<n><sg>$
^گذاشت/گذاشت<vblex><past><p3><sg>$

1- The postposition 'rā' in Persian was matched with 'the' as definite determiner in English. Based on the Apertium system, two lemmas should have the same or similar morphology to be equivalent with each other [17].

The example (2) has six words and every word has its own morphological attributes, which are defined in Persian and English morphological dictionaries. Here, the tags ‘<np>’ and ‘<unk>’ mean ‘proper noun’ and ‘unknown’[1] respectively. The tag ‘<n>’, stands for ‘noun’ and the tag ‘<sg>’ is used for ‘singular’ number. The tags ‘<post>’[2] and ‘<pr>’ are the short forms for postposition and preposition respectively. Then, the tags used for the verb "گذاشت" ‘put’ are: ‘<vblex>’, ‘<past>’, ‘<p3>’, and ‘<sg>’. They show that "گذاشت" is a main verb, past tense, third person, and its number is singular.

2.3 POS tagger is based on the first order of Hidden Markov Model (HMM) [18]. Every SL source form need to be analyzed to one target language (TL) lexical form. For ambiguous words in SL morphological dictionary which are analyzed to more than one TL lexical form, POS tagger module selects one of these lexical forms and sometimes constraint grammar (CG) rules[3] [19] are applied before the final results of POS tagger. For example, there is a morphological ambiguity for the noun "تابع" ‘function’ in the noun phrase "تابع اعداد" ‘numbers function’. It was shown based on the morphological analyser module as below:

3- ^تابع/تابع<adj>/تابع<n><sg>$ ^اعداد/عدد<n><pl>$

The word "تابع" has two POS tags: ‘<adj>’ and ‘<n>’ which stand for the adjective ‘subordinate’ and the noun ‘function’ respectively. POS tagger selects one tag (here ‘<adj>’) and its translation can be as follows:

4- ^تابع<adj>$ ^عدد<n><pl>$

‘subordinate numbers’

Then, the following CG rule disambiguates the word "تابع" in (3):

5- SELECT N IF (1 (N))

1- This tag was used for some proper nouns which were not included in the morphological dictionary of English (*apertium-eng.eng.dix*).

2- The symbols are used from http://wiki.apertium.org/wiki/List_of_symbols. But some tags which were specifically used for Persian listed in List of Symbols.

3. In Apertium, the file for CG rules is called “*apertium-pes-eng.pes-eng.rlx*”.

Rule (5) shows that ‘N’ for noun should be selected in phrases followed by another noun. So the phrase "subordinate numbers" can be used. Because based on Persian phrase structure, an adjective occurs before a noun. So it should be translated as a noun instead of an adjective and its translation changes to ‘function numbers’.

2.4 Lexical transfer module, based on a bilingual dictionary, corresponds SL lexical forms to a TL lexical form from the bilingual dictionary. It reads each LF of the SL and delivers the corresponding TL lexical form. The dictionary has a single equivalent for each SL lexical form. Ambiguous words in an SL morphological dictionary are analyzed into more than one TL lexical forms [20].

2.5 Lexical selection rules can be written by hand (in our example *apertium-pes-eng.pes-eng.lrx*) to solve the problem of lexical ambiguity with a given context. If there are more than one TL equivalents for one SL lexical form, lexical selection module selects one of them. For example, the Persian gerund "ترک کردن" is translated with a default translation of ‘stop’, but to translate it with another meaning, ‘leave’, the following rule needed to be written:

```
6- <rule>
        <match lemma="خانه" tags="n.*[1]"><select
lemma="house" tags="n.*"/></match>

    <or>
        <match lemma="را" tags="det.def"/>
        <match lemma="ترک کرد" tags="vblex.*"/>
    </or>
</rule>
```

Rule (6) shows that the Persian verb "ترک کرد" is translated as ‘left’ if it is followed by the noun "خانه" ‘house’ and the postposition "را".

1. The symbol ‘*’ means any tag can be placed after the preceding tag(s).

Now, the sentence "رضا خانه را ترک کرد" is translated to ‘Reza left the house’.

2.6 Structural transfer module uses transfer rules to transform SL sentence or phrase to TL. This module covers syntactic processing. To process, it uses transfer rules, which transform lexical forms sequences to another sequence of target language. Structural transfer works in three steps. First of all is “chunker” level[1] (here *apertium-pes-eng.pes-eng.t1x*), which divides source sentence to chunks. At the second level, namely in “interchunk” (*apertium-pes-eng.pes-eng.t2x*), it does rearrangement of phrases. For example, in the following sentence:

7- "زیبا با ناراحتی حرف زد" ‘Ziba spoke sadly’

There are three chunks resulted from the chunker level including <SN> for the noun ‘Ziba’, <SV> for the past tense verb ‘spoke’, and <SADV> for the adverb ‘sadly’ needed to be reordered based on the TL (as English) syntactic patterns:

8-^noun<SN>{^Ziba<np><unk>$}$
^simple_SV<SV><p3><past>{^speak<vblex><past>$}$
^adv<SADV><adv>{^sadly<2>$}$

At final level, “postchunk”, (*apertium-pes-eng.pes-eng.t3x*), it does some clean-up by deleting unnecessary tags [13], like the following pattern:

9- ^Ziba<np><unk>$ ^speak<vblex><past>$ ^sadly<adv>$.

2.7 Morphological generator generates a corresponding sequence of target language surface forms [21]. For instance, the noun "خانه" ‘house’ is morphologically analyzed into the English word ‘house’ like the below example

10- خانه [] house []

2.8 Post-generator takes care of some minor orthographical operations in the TL lexical forms [16].

1. The chunker rules’ structures are discussed in the next section.

2.9 Reformattor backs the formatting tags to the text to be as it was appeared first [22], as the following example:

11- pes []اسب[].

3. Translation Problems of Clefted Constituents in the Apertium System

In Persian translation, there are some general problems and difficulties due to Persian syntactic structures. Although Persian uses an SOV word order, there are several frequent exceptions in word order, caused by processes such as topicalization, dislocation, clefting, pseudoclefting, and scrambling [23] that mainly result in high structural complexity [24]. This research tries to examine and solve the translation problems of the clefted constituents by adding new chunker rules at *apertium-pes-eng.pes-eng.t1x* file[1].

3.1 Clefting Process

According to Mahootian [23], "clefting in Persian moves the focus element from its unmarked position to the beginning of the sentence. The focused element of the sentence is followed by a verb, usually a copula and the relative pronoun 'ke'". "In English a clefting construction (It-cleft) is a syntactically bi-clausal construction, it comprises a main or matrix clause containing a copula verb and a subordinate clause. The cleft clause is often introduced by either a relative pronoun or the complementizer 'that'. These elements can be omitted if the element 'missing' from the cleft clause is not the subject" [25].

It is possible to cleft subjects, direct objects, indirect objects[2], and adverbs of time [26].

1. It can be downloaded from https://svn.code.sf.net/p/apertium/svn/incubator/apertium-pes-eng.
2. Also, the objects of preposition can be clefted.

3.1.1 Clefting of Subject

The subject "رضا" in Sentence 12 was clefted in Sentence 13 as the following:

12- "دیروز رضا شیشه را با سنگ شکست." 'Yesterday Reza broke the widow with stone'

13- "رضا بود که دیروز شیشه را با سنگ شکست." 'It was Reza who broke the window with stone yesterday'.

Sentence 13 was translated initially in the Apertium system as 'Reza was who yesterday the glass with stone broke'. To generate it correctly, the chunker rules were added as follows:

14- رضا<np>/Reza<np>بود<vbser><past>/be<vbser><past> که<rel>/who<rel>

Rule 36[1]↓

it<prn><subj> was<vbser><past> Reza<np> who<rel>

دیروز<adv>/yesterday<adv> شیشه<n>/glass<n>
را<det><def>/the<det><def> با<pr>/with<pr> سنگ<n>/stone<n>
شکست<vblex><past>/break<vblex><past>

Rule 54↓

break<vblex><past> the<det><def> glass<n> with<pr> stone<n> yesterday<adv>

To cleft the subject "رضا" as <np> to the beginning of Sentence 12, two rules were applied. Rule 36 generated the proper noun "رضا", the auxiliary verb "بود", and the relative pronoun "که" to its English patterns: the subjective pronoun (as 'It'), the past tense auxiliary verb '<vbser>' "was", the proper noun "Reza", and the relative pronoun 'who'. Among all existing relative pronouns in dictionaries like 'that', 'which', and 'who', the used tagger selected the proper relative pronoun (here who) referring to the preceding noun. Then, Rule 54 generated the rest of sentence as the adverb of time "دیروز", the noun "شیشه", the definite determiner "را", the preposition "با", the noun "سنگ", and the past tense verb "شیشه" into the past tense 'broke', the

1 . The number of rules are based on their orders in *apertium-pes-eng.pes-eng.t1x*.

definite noun ‘the glass’, the prepositional phrase ‘with stone’, and the adverb of time ‘yesterday’. Then, final Apertium translation was ‘it[1] was Reza who broke the glass with stone yesterday.’

3.1.2. Clefting of Direct Object

The direct object "شیشه را" in Sentence 12 was clefted like the below Sentence:

15- "*شیشه* بود که دیروز رضا آن را با سنگ شکست." ‘It was window that Reza broke with stone yesterday’.

The Apertium system primarily translated it as ‘glass was that yesterday Reza it with stone broke’. To produce more accurate translation, adding some chunker rules were required:

16-

شیشه<n>/glass<n> بود<vbser><past>/be<vbser><past>
که<rel>/that<rel>

Rule 36↓

it<prn><subj> be<vbser><past> glass<n> that<rel>

دیروز<adv>/yesterday<adv> رضا<np>/Reza<np>
آن<prn><subj><p3><nt><sg>/prpers<prn><subj><p3><nt><sg>
را<det><def>/the<det><def> با<pr>/with<pr> سنگ<n>/stone<n>
شکست<vblex><past>/break<vblex><past>

Rule 52↓

Reza<np> break<vblex><past> it<prn><obj> with<pr> stone<n>

The direct object "شیشه" was clefted to the initial position of Sentence 15 using Rules 36 and 52. Rule 36 was applied the same as its usage in Sentence 13 which changed the Persian pattern: the noun "شیشه", the auxiliary verb "بود", and the relative pronoun "که" into the English pattern as the subjective pronoun ‘it’, the verb ‘was’, the noun ‘glass’, and the relative pronoun ‘that’. Then, Rule 52 translated the adverb "دیروز", the proper noun "رضا", the pronoun "آن", the definite determiner "را", the preposition "با", the noun "سنگ", and the past

1 . Currently the system is not sensitive to capital letters at the beginning of a sentence.

tense verb "شکست" into the proper noun ‘Reza’, the past tense verb ‘broke’, the objective pronoun ‘it’, the preposition ‘with’, the noun ‘stone’, and the adverb of time ‘yesterday’. Finally, Sentence 15 was generated as ‘it was glass that Reza broke it with stone yesterday’.

3.1.3 Clefting of Indirect Object

The indirect object "به زهره" in Sentence 17 was clefted in Sentence 18 as follows:

17- "سیما ساعت را به زهره داد" ‘Sima gave the watch to Zohreh’

18- "به زهره بود که سیما ساعت را داد" ‘It was Zohreh that Sima gave the watch to’.

Before adding new chunker rules, Sentence 17 was translated as ‘to Zohreh was that Sima the watch gave’ in the developed system. To remove translation errors, we needed to write new chunker rules as below:

19-

به<pr>/to<pr> زهره<np>

Zohreh<np>بود<vbser><past>/be<vbser><past> که<rel>/that<rel>

Rule 37↓

it<prn><subj> was<vbser><past> Zohreh<np> that<rel>

سیما<np>/Sima<np> ساعت <n>/watch<n>

را<det><def>/the<det><def> داد<vblex><past>/give<vblex><past>

Rule 44↓

Sima<np> give<vblex><past> the<det><def> watch<n> to<pr>

In clefting indirect object "به زهره" to the initial position of Sentence 16, two rules were written. Rule 37 changed the Persian patterns the preposition "به", the proper noun "زهره", and the verb ‘to be’ "بود" into the subjective pronoun ‘it’, the past tense verb ‘was’, the proper noun ‘Zohreh’, and the relative pronoun ‘that’. Rule 44 generated the proper noun "سیما", the noun "ساعت", the definite determiner "را", and the past tense verb "داد" to the proper noun ‘Sima’, the past tense verb ‘gave’, the definite noun ‘the watch’, and

the preposition ‘to’. Then Sentence 18 was produced as ‘it was Zohreh that Sima gave the watch to’.

3.1.4 Clefting of Object of Preposition

The object of preposition "توی باغ" in Sentence 20 was clefted in Sentence 21 as follows:

20- "همدیگر را توی باغ دیدیم" ‘We saw each other in the garden’

21- "توی باغ بود که همدیگر را دیدم" ‘It was in the garden that we saw each other’.

Firstly, Sentence 21 was generated as ‘in garden were that each other saw’ in the Apertium system. To generate it correctly, adding some chunker rules were needed:

22-

توی<pr>/in <pr> باغ<n>/garden<n>
بود<vbser><past>/be<vbser><past> که<rel>/that<rel>

Rule 37

it<prn><subj> was<vbser><past> garden<n> that<rel>

همدیگر<prn>/each other<prn> را<det<def><sp>/the<det><def><sp>
دید<vblex><past>/saw<vblex><past>

Rule 55

we<prn><p1><pl><mf> saw<vblex><past> each other<prn>

In Sentence 21, the object of preposition "توی باغ" was clefted to its initial position. Rule 37 was applied to translate the first part of Sentence 21 the same as what was done in Sentence 18. Since the preposition ‘in’ was required before the noun ‘garden’, it was not translated correctly. So writing new chunker rule could not possible because the same SL pattern was written and writing the same rule could block the path working Rule 37. Therefore, Rule 55 produced the pattern: the subjective pronoun ‘we’, the past tense verb ‘saw’, and the reciprocal pronoun ‘each other’. Finally, Sentence 21 was translated to ‘it was *garden that we saw each other.’

3.1.5 Clefting of Adverb of Time

The adverb of time "دیروز" was clefted in Sentence 23 as the following:

23- "دیروز بود که رضا شیشه را با سنگ شکست" 'It was yesterday that Reza broke the glass with stone'.

Sentence 23 was firstly translated as "Yesterday was that Reza the glass with stone broke" in Apertium. To improve translation performance, new chunker rules was added to the system as follows:

24-

دیروز<adv>/yesterday<adv>بود<vbser><past>/be<vbser><past>
که<rel>/that<rel>
Rule 35
it<prn><subj> be<vbser><past> yesterday<adv> that<rel>

رضا<n>/Reza<n>شیشه<n>/glass<n>را<det><def>/the<det><def>
با<pr>/with<pr>سنگ<n>/stone<n>
شکست<vblex><past>/break<vblex><past>
Rule 57
Reza<np> break<vblex><past> the<det><def> glass<n> with<pr> stone<n>

The adverb of time "دیروز" was clefted to the initial position of Sentence 23 by two rules. Rule 35 translated the Persian pattern the adverb of time "دیروز", the auxiliary verb "بود", and the relative pronoun "که" to the subjective pronoun 'it', the verb 'was', and the relative pronoun 'that'. Then, Rule 57 translated "رضا"to 'Reza' as proper noun, the past tense verb "شکست" to 'broke', the noun phrase "شیشه را" to the definite noun 'the glass', and the prepositional phrase "با سنگ" to 'with stone'. The final Apertium translation of Sentence 23 was 'it was yesterday that Reza broke the glass with stone'.

4. Results and Discussion

To investigate the syntactic translation problems of clefted constituents, 100 sentences were extracted from the Persian spoken and written data. They were stored in destination file named

clefting_construction.txt. Then, this was run in the Apertium system and Google Translate. Some of the words were not translated in the Apertium system. All non-translated words were added to the Apertium (monolingual and bilingual) dictionaries. Besides, the output was post-processed by a human translator. Both Apertium and Google translate outputs were computed by WER and PER evaluation metrics. WER is computed as the minimum number of substitution, insertion, and deletion operations that have to be performed to convert the generated translation into the reference translation. A shortcoming of the WER is the fact that it requires a perfect word order. In order to overcome this problem, PER of two sentences were compared without taking the word order into account [27]. Note that WER and PER calculate translation errors, the less they are, the better translation performance. The evaluation results of two machine translated hypotheses are represented in Table 2:

Table 2: Comparative evaluation results of the clefted constituents for post-processing task in the Apertium system and Google Translate

Machine Translation System	WER	PER
apertium-pes-eng	59.08 %	48.93 %
Google Translate	62.73 %	50.66 %

According to Table 2, the WER and PER scores decreased more in the Apertium system than those of Google translate. The difference between WER scores (-3.65%) in both systems was higher than that of PER scores (-1.73%). It indicated that adding chunker rules outperformed the Apertium system by translating clefted consituents.

5. Conclusions

Based on the Apertium modules, translation performance could be improved by working on each module in translating from Persian language to English. This paper attempted to solve the translation problems of clefted constituents by adding chunker rules at structural transfer module. The added rules helped to generate more accurate

translation which was closer to the reference text than those of Google translate. The results were proved by measuring the evaluation metrics such as WER and PER.

Acknowledgements

We thank Mikel. L Forcada and Víctor Manuel Sánchez Cartagena for their useful comments. The anonymous reviewer's suggestions to improve the paper is greatly appreciated.

List of Symbols

<adj>	Adjective
<adv>	Adverb
<det>	Determiner
<def>	Definite
<mf>	Masculine/Feminine
<n>	Noun
<np>	Proper noun
<nt>	Neuter
<obj>	Object
<p1>	First person
<p3>	Third person
<past>	Past tense
<pl>	Plural
<prn>	Pronoun
<post>	Postposition
<prpers>	Personal pronoun
<rel>	Relative
<sp>	Singular/Plural
<subj>	Subject
<unk>	Unknown
<vblex>	Main verb
<vbser>	Auxiliary verb

References

[1] Tripathi, S., Sarkhel, J. K., 2010. "Approaches to machine translation". *Annals of Library and Information Studies*, 57, pp. 388-393.

[2] Och, F. J., and Ney, H., 2004. "The alignment template approach to statistical machine translation". *Computational Linguistics*, 30(4), pp. 417–449.

[3] Mohaghegh, M., Sarrafzadeh, A., 2012. "A hierarchical phrase-based model for English-Persian statistical machine translation". *International Conference on Innovations in Information Technology (IIT)*, pp. 205-208

[4] Soheili, A., 1989. Is Persian a Pro-drop Language? University of Maryland, USA.

[5] Karimi, S., 1989. "Aspects of Persian Syntax, and the Theory of Government". PhD Dissertation. University of Washington. See also URL https://trove.nla.gov.au/version/9688230.

[6] Darzi, A., 1996. "Word Order, Np-Movement, and Opacity Conditions in Persian". PhD Dissertation, University of Illinois, Urbana. See also URL http://hdl.handle.net/2142/20523.

[7] Ahangar, A., 2001. "Explanation of the Place and nature of the Verb Complement Clause in the Persian Language". *Journal of Humanities, University of Sistan and Baluchestan*, 18, pp. 65-100.

[8] Cook, V. J., 1985. "Universal Grammar and Second Language Learning". *Applied Linguistics*, 6, pp. 2-18.

[9] Simpson, A., 2005. "Pro-drop Patterns and Analyticity", *LSA 222 Syntactic Analyticity*.

[10] Dalili, V. M., 2009. "Agreement (AGR) and the Pro-drop/Non-pro-drop Variation: A Meta-analysis of GB and MP accounts". *Philologie im Netz*, 49, pp. 84-102.

[11] Dabir-Moghaddam, M., 2001. "Word Order Typology of Iranian Languages". *Journal of Humanities*, 8(2), pp. 17–24.

[12] Sánchez –Cartagena, V. M., Sánchez-Martínez, F., Pe' rez-Ortiz., J. A., 2011. "The Universitat d'Alacant hybrid machine translation system for WMT 2011", *Association for Computational*

Linguistics, Proceedings of the 6th Workshop on Statistical Machine Translation. Edinburgh, pp. 457–463. See also URL https://dl.acm.org/citation.cfm?id=2133023.

[13] Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F. M., 2011. "Apertium: a free/open-source platform for rule-based machine translation". *Machine translation*, 25(2), pp. 127–144.

[14] Supreme Council of Information and Communication Technology., 2013. *Mizan English-Persian Parallel Corpus*.Tehran, I.R. Iran. See also URL http://dadegan.ir/catalog/mizan.

[15] Sundetova, A., Forcada, M. L., Tyers, F., 2015. "A free/open-source machine translation system for English to Kazakh". *3rd International Conference on Computer Processing in Turkic Languages.*

[16] Karibayeva, A., 2015. "Lexical selection rules for Kazakh-to-English machine translation in the free/open-source platform Apertium". *3rd International Conference on Computer Processing in Turkic Languages.* See also URL http: turklang.antat.ru.

[17] Forcada, M. L., Bonev, B. I., Ortiz-Rojas, S., Ortiz, J. A. P., Sánchez. G. R., Sánchez-Martínez, F., Armentano-Plller. C., Montava, M. A., Tyerz, F. M., 2010. Documentation of the Open-Source Shallow-Transfer Mschine translation Platform Apertium. Departament de Llenguatges i Sistemes Inform`atics Universitat d'Alacant.

[18] Rabiner, L., 1989. "A tutorial on hidden Markov models and selected applications in speech recognition". In *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE),* 77(2), pp. 257–28.

[19] Karlsson, F., 1990. "Constraint Grammar as a Framework for Parsing Running Text". *In H. Karlgren, ed., Proceedings of the*

13th Conference on Computational Linguistics. Helsinki: Finland, 3, pp. 168–173.

[20] Tyers, F. M., Sánchez-Martínez, F., Forcada, M. L., 2012. "Flexible finite-state lexical selection for rule-based machine translation". In *Proceedings of the 16th EAMT Conference*, Trento, Italy. See also URL altea.dlsi.ua.es.

[21] Peradin, H., Tyers, F., 2012. "A rule-based machine translation system from Serbo-Croatian to Macedonian". In *Proceedings of a workshop on Free/open source machine translation*. Gothenburg, pp. 41-55. See also URL *www.mt-archive.info/FreeRBMT-2012-Peradin.pdf.*

[22] Trosterud, T., Unhammer, K. B., 2012. "Evaluation North Sámi to Norwegian Assimilation RBMT". In *Proceedings of a workshop on Free/open source machine translation*, pp. 1-13. See also URL *www.mt-archive.info/FreeRBMT-2012-Trosterud.pdf.*

[23] Mahootian, Sh., 1997. Persian. (Descriptive Grammars). London: Routledge.

[24] Saedi, Ch., Shamsfard, M., Motazedi, Y., 2009. "Automatic translation between English and Persian texts". *In Proceedings of the Third Workshop on Computational Linguistics*. See also URL www.mt-archive.info.

[25] Pavey, E., 2004. "The English IT-cleft construction: A Role and Reference Grammar Analysis". MA Thesis, University of Sussex.

[26] Karimi, S., 2005. A Minimalist Approach to Scrambling: Evidence from Persian. Berlin: Mounton de Gruyter.

[27] Popović, M., Ney, H., 2007. "Word error rates: Decomposition over POS classes and applications for error analysis". In *Proceedings of Workshop on Association for Computational Linguistics (ACL)*, pp. 48-55. See also URL http://www.aclweb.org/anthology/W07-0707.

Annotated Corpora Beneath Application Programming Interface

Amir H. Tavassolinia*

Abstract

Evidence of language mechanism of the human brain is being collected in forms of storing electronic naturally occurring texts called corpora. Corpus Linguistics tends to reveal the rules of languages to implement by machines that use corpus dependent Natural Language Processing systems. Persian NLP systems are functioning well however, few unlimited Persian corpora were compiled to test them to be improved. In the technology era, texts are produced by Persians in websites and applications of any kind to imply meaningful sentiments. The applied linguist in this study used python to get authorized to a social media API to examine the amount and lexical density of natural Persian streaming in the API. The corpus contained half a million words with tagged information about real-time users. By accessing such APIs and scraping websites, the unlimited Persian Today Corpus will start compiling for NLP technologies.

Keywords: Corpus Linguistics, Computational Linguistics, Natural Language Processing, Data Mining, Machine Learning, Python

1. Introduction

Compiling texts in order to be analyzed and reveal the hidden aspects of context can be traced back to the thirteenth century when scholars gathered a team including Anthony of Padua to search page by page of the Bible in order to index the words which resulted in the idea that Bible was not just a series of texts but a harmonious divine implied in the context [1]. Producing dictionaries of English had also required paper corpora and lots of time. Like Samuel Johnson's first published dictionary of English in 1755. In the 1950's, the era of

* Islamic Azad University, Shiraz Branch, Department of English; amirtvsl@outlook.com

American Structuralism supported by Harris, Fries and others, collecting data became more important than before. Moreover by the emergence of computers, this collection of data became much easier and faster [1].The first computer-generated concordances which are listings of words with examples in context were conducted in the 1950's and punched-card technology was used for the storage with the ability to process sixty thousand words a day [1]. The word corpus was first cited in 1956 in an article by William Sidney Allen and its meaning became closer to what it means today[1]. Since the 1960s, computer-based corpus compilation was conducted electronically to build dictionaries and study languages. One of the first electronic corpora compiled by Kucera and Francis was the Brown Corpus [2] which is still one of the largest corpora available.

Storing naturally occurring language electronically is concerned with Corpus linguistics which its studies are core principles in Computational Linguistics and in fact the improvements of technology would not have been made without it. Natural language processing programs came to existence by analyzing texts to decode the human language mechanism in order to find logic in language, in other words, NLP is trying to gain the ability to implement the human use of language and to understand language by reading texts or interacting with humans. The emergence of NLP dates back to 1950s in order to solve Machine Translation issues [3]. The theory of translating texts with machines started from the seventeenth century and different methods were used to overcome the issues concerning the quality and accuracy of translation products. By emergence of NLP methods, MT started improving and the importance of NLP was increased.

The improvement of this technology on implementing English goes to the recent accomplishment in being slightly better than humans in reading texts by winning for the first time in a reading comprehension contest on the first month of 2018, by running Alibaba and Microsoft's Artificial Intelligence algorithms which developed by NLP and Machine Learning models [4]. NLP models are available in any language and they are corpus dependent which clarifies the

valuable role of corpora in building language related ML and AI systems.

Persian NLP is also improving by scholars compiling corpora as datasets and computer scientists developing algorithms to understand Persian texts better. The abundance of various Persian corpora would weigh down the improvements of such systems. Unfortunately, there are a few Persian corpora compiled in natural Persian available. The existing Persian corpora like Tehran Monolingual and parallel Corpus [5], Bijankhan Corpus [6] and others, usually contain the formal use of language written in news articles by groups of journalists. However, Persian NLP needs more natural corpora in various categories in order to reach other languages' potentials in decoding the contexts.

The invention of the internet made virtual connections where any aspect of humanities could take place without any boundaries. Electronic texts started producing eventually by humans in various forms such as websites, social media and numerous applications of any kind. The amount of meaningful texts [7] is increasing and getting access to the real-time naturally occurring texts produced in worldwide web and compiling them as corpora to improve NLP, would be crucial.

Application Programming Interface is the collection of functions and procedures that an application needs to be developed which can have the accessibility to the data of an operating system. This means that when an application is built upon an API, the data streaming in that application is accessible through it. Quite a few applications were built upon such infrastructure with their developers hoping that people participate in the app and the stream of its API would be up and running, but this doesn't happen most of the time for all the applications [7]. Recently people use few applications to share their feelings, hobbies or anything else that can be fit as a social activity. Luckily most of these populated applications provide open access to their API for the academic purpose of data analysis [8]. These APIs contain lots of information in any form but the most important data

available in APIs concerning linguistic matters are the texts. This kind of texts can be about anything but the only thing that is similar in all texts is their reliability as being meaningful and informant because they are produced by actual persons with intentions of sharing their thought [7]. Considering Corpus Linguistics aspects, available texts in the worldwide web are good sources of corpora [9]. Such texts can be accessed through APIs of social media or by scraping the web pages. Scraping web pages are mining the whole data available on a website [10]. Applications or websites that provide open access to their APIs are for instance Github, Amazon, Yahoo, Instagram, Twitter, Youtube, Linkedin and a lot more [11].

API systems work in a way that an API call should be made to start data parsing and formatting to produce the results in the graphic user interface (GUI) hence when one wants to mix two or more different results together, she can save the results in an array while getting information from other sources. By doing some business logic in algorithm forms then, the returns get together as a whole [12], like shown in figure 1. This can help narrow down the result to specifics, for instance, if one wants to gather only Chinese texts coming from Europe and places there which is raining, by doing what was mentioned the result would be all the Chinese texts from currently raining cities in Europe with just one API call. This method is called Data Mining for a specific purpose [13].

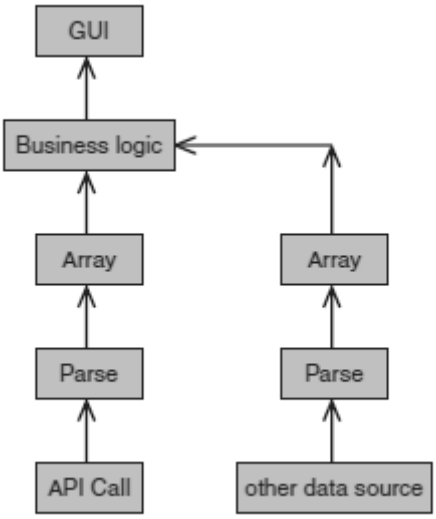


Figure 1. Example of combining two data streams [12]

In recent times the amount of data is larger than one can imagine. Around early 2014, the online technology dictionary Webopedia stated that in 2011, 1.8 trillion gigabytes of data were created [7]. This amount is still increasing by the improvements in technology and ease of access to smartphones. These data are not just to be made for no reason. In the technology era, scholars use data available to extract desired information in order to reach qualitative information from quantitative one. In Corpus Linguistics and NLP, the aim is to reveal how humans think and how the use of language really works [3]. To be able to do so, a large amount of data is needed because the use of language is arbitrary and anyone can use language in their own ways and that is something hard to track. However based on Sapir-Whorf theory developed by Sapir and Benjamin Lee Whorf, that states structure of a language determines the modes of thought and behavior characteristics of culture which is spoken [14], and the law of large numbers proved by James Bernoulli which states that the large population of unpredictable players collectively behaves in a predictable fashion [15], make it clear that language use can get close

to being logical and machines can decode language ultimately with the help of large corpora [3].

Machines nowadays are doing a good job interacting with humans but they are yet to be improved. One concern here is that decoding a language like Persian is a lot more difficult than English and there is almost no Persian human-machine interaction like other languages used in virtual assistants like Microsoft's Cortana and Apple's Siri. Reaching the technology requires lots of different corpora and without any, no virtual assistant could come into existence. This raises the importance of Corpus Linguistics that is the basic requirements of technologies such as Deep Learning, Computational Linguistics, Natural language Processing, Machine Translation and also Artificial Intelligence. These technologies aim to reach the ability of machines to understand and behave like humans by themselves without the need to be pre-programmed [16].

The worldwide net provides the access to a large amount of data which is multilingual and meaningful with having intentions of producing. Fortunately, Persians got adapted to the revolution of Data and there are lots of users using Unicode Persian in any interface from broadcasting to personal interactions. This could be quite helpful in processing Persian by filtering and compiling this kind of electronic uses of Persian by accessing the APIs or scraping the websites because no matter what the online text contains, it is pure Persian.

When a corpus is a compilation of numerous language users, the reliability of it would increase because it is not limited to mechanisms of the language of a group of people like journalists. Concerning the style of the writer which is the thumbprint [17] of the writer that uses some words more, and there are words that a person does not use subconsciously, these types of corpora is not natural enough. In order to reach closer to the ultimate intention of NLP which is the implementation of the human arbitrary use of language [18], datasets that are not limited in the styles of authors is needed to be tested and by debugging NLP programs better which is a permanent procedure

after emergence of any program, the performance of any NLP systems would get closer to being flawless.

NLP programs are programs that can pull texts apart and understand the words as each and the texts as a whole [19], these procedures are gaining knowledge power by Deep Learning technologies and improving day by day. A text of any language can be tokenized by the help of modules like NLTK for English and HAZM for Persian. These modules are just instances of all toolkits developed for NLP and there are more available to be used. They can be imported in python to analyze texts. Features of such programs help machines to be able to understand the text and respond which this ability is gaining more improvements depending on languages of the texts [19].

Almost all of the NLP programs can be created and imported into python. Python is an object-oriented programming language which is very powerful and easy to learn comparing to other programming languages. It is a cross-platform language which means it can be read on any system and it does not need any translation to be used [20]. Most of NLP libraries are available to be imported into python to analyze data. The use of this programming language is not limited in fields of study and most of the successful corporations like Google are using python to devise their machine learning systems, for instance, the machine learning system which Google uses to suggest its users the related information based on their previous searches is using python[16]. The power of python is limitless and using this language in addition to the compilation of more corpora as possible, leads to better understanding of natural language by machines and better performances of language-related AI [16].

In the followings, the applied linguist curiously delved into an API to see whether it would be possible to access real-time large amounts of data written in Persian to be compiled or not and provide the types of analysis that could be done with this type of corpus.

The API selected was the Twitter API to increase the possibility of existing Persian text by filtering the code to Persian which is introduced in the API's text input as 'fa'. In order to be able to access Twitter API, getting authorization is needed and one should already have an account, then create a twitter app to be able to run twitter by a program on the user's behalf. This passive application is available on Twitter development's web page [21]

By signing up there, an independent twitter app with the specific access token, consumer key, and secret would be created. Installing tweepy via pip install command makes it possible to be imported into python shell. The authorization to have access to the stream of data in Twitter API is now fulfilled where massive data is on stream. Then one would write the code to get desired information from the stream [12].

Each tweet's information is not just about the text itself, it contains information about the geographical emergence of data, personal information, account information and a lot more. Hence the texts contain additional tagged information just like annotated corpora and the features of Data Analysis are possible with this type of corpus.

After the authorization, only three library dependencies would be enough to be imported into python, that is tweepy, Authhandler and StreamListener. Next step would be to classify the stream listener for python and to define what is desired to be extracted from the stream. Twitter stream could be filtered by anything from splitting the results, to language of the text. Hence the first API call was made by filtration of the date, the keyword ''ایران'' in 'fa' and splitting the results by removing other modification of data such as public personal information, number of posts, followers and many more which is tagged to the data. It is also worth mentioning that there are 150 API calls available per day for any program or bot to make for free [12].

After saving and running the program, the stream started running, parsing, formatting and saving data automatically. The compiling of

desired text ran for 24 hours and it was still mining but the programmed were killed intentionally to make the scaling out of one-day mining equality to how much desired data considered here as a corpus. The tagged corpus made by this methodology, contained the name of the Persian author of the text, the date, and the location if applied. The desired format of the corpus was selected as a CSV file to be easier to do some data analysis with. The corpus should be decoded to UTF-8 to be seen as Unicode.

It contained more than half a million words in Persian, from about 8000 individual tweets and having the hundred percent of the frequency of the word "Iran" in just a day. The type-token ratio of the corpus was high which means this corpus is linguistically rich by having different words more than repeated ones.

By having this type of corpora which is tagged by the way obtained automatically not manually, various types of analysis can be done. For instance, it is possible to perform a geographical analysis to show the distribution of data which is texts in this case, with plotting the location tag of the corpus with matplotlib or Basemap in python [22]. This clarifies the population places the discourse was taken which also can go further than linguistics matters to help the government control the virtual world too, to gain more power and knowledge of data.

The other type of analyzing the corpus is sentiment analysis, one can delve into the stream of an API to compile the required corpus by narrowing down the search based on needs. Then by the help of natural language processing toolkits available in languages including Persian, thanks to their developers, the opinion of each individual on a specific title can be analyzed to make a generalization out of the data, in other words, make the data mean something [13]. For example at the basic level of sentiment analysis, if a machine or human, wants to know the feedback of people about the filtered keyword, by the kinds of words they have used, it can use NLP modules to chink unnecessary words (stop words) and creates a bag of word without caring about the grammar rules because grammar is

just the structure of being correct in using the words and it is not an important characteristics in sentiment analysis [13]. NLP modules can clean the texts by removing stop words that have no meaning and were used as linguistic tools to relate discourse or words together but by themselves alone, they don't have a meaning [19]. Clearing texts from stop words are important especially when data analysis is going to be taken. The imported modules filter the text by pre-defined stop words of the used language compiled as corpus and return the result without the stop words for the corpus to contain only definitive and informant words. The list of stop words of languages can also be seen through printing the pre-defined stop words of a language in the python shell. By creating the desired bag of words and with the help of NLP toolkits, it is possible to get the frequency of words and classify the most common words as the binary classification of being good or bad.

The probable outcome would be the real-time sentiment or opinion of people who have written down their thoughts electronically about the filtered word. In this case, the word would be 'ایران'. Comparing the frequent words with being good or bad are done in python as algorithms which are also available in Persian word sense classification dataset. The returned sentiment of people about Iran is not illustrated here because the researcher does not want to generalize the data in just one day with one API call. The aim was just to elaborate the possibility of gathering valuable texts to be compiled as a corpus to enlarge NLP datasets.

2. Nomenclatures

AI Artificial Intelligence

API Application Programming Interface

CSV Comma Separated Value

GUI Graphical User Interface

ML Machine Learning

MT Machine Translation

NLP Natural Language Processing

3. Results and Discussion

NLP tools nowadays can understand what the text is all about, what are important parts of a text and they are able to paraphrase a large text into a small one by implying valuable parts by article spinning and Text Mining. NLP tools at first do some preprocessing to texts to save processing time which will be briefly explained here.

Tokenizing and clearing stop words mentioned before, are instances of such. It also includes stemming, which is to take the root stem of the word in order to get the exact meaning of words and to avoid distractions like plural signifiers or different changes of word forms depending on their role in sentences. Each word in chunks can be defined as their linguistic role in sentences by named entity recognition algorithms which modify the words as nouns, verbs, adverbs, etc. Tagging the part of speech is labeling every single word which is not always accurate because of the type of writing by people may mess up with part of speech tagging because anyone relates words by different strategies of language use which is sometimes hard to understand even by humans rather than machines [19]. Hence the periodical changes of words like semantic and connotation changes of words in languages based on their use throughout the time will change a word's meaning from considering the good word to a bad one.

The compiled corpus was processed and the analysis will be provided in the following. The total amount of words containing in the corpus were first examined and then the repeated words were omitted to examine the number of types. By dividing types by tokens or the total amount of words and multiplying the result to a hundred, the lexical density of the corpus would be examined as shown in the table below. The word "ایران" contains in any sentence because of being the keyword in API call made to compile the corpus and it is the most frequent word containing there.

Table 1. Lexical density of the corpus

Number	Token	Type	Lexical density
1	511,963	389,574	76%

4. Conclusions

English corpora are narrowed down in any type of categories from ancient works of art to daily uses of language in worldwide web even there are chat corpora available in English which means each individual's language use counts in Data Science and people are to feed the machines [7]. English corpora nowadays contain billions of words each and most of them are available online with tools to analyze them on the web rather than downloading it and run it on another corpus tools program. The only thing that can be tested by NLP systems to evaluate them, is different kinds of corpora. Hence the more corpora compiled to help scholars reveal natural language, the better implementation of natural language and accuracy of algorithms. The whole available Persian corpora will not get close to having 1 billion words while the BNC corpus contains 520 billion words let alone.

Persian NLP can be improved to increase machine/ deep learning systems' accuracy by updating the existing corpora and compiling more unlimited ones. This procedure of updating the datasets or corpora is also possible with python to make API calls daily and enrich the corpora and save them automatically. Updating corpora can also enrich Wordnets which are more than just stored texts. They include synonyms and related words with the ability to estimate semantic similarity as a percentage to show how close two words are in their meanings. The available Persian Wordnet also depends on different corpora with new words to be imported and be updated by filtering the words and comparing two corpora. This can also make processing semantic similarity more reliable because more words are getting imported into it and using the Wordnet in plagiarism detection systems can be more powerful because such systems tend to dig into

words and change them with synonyms and compare the words to detect plagiarism.

Persian applied linguists and computer scientists would better start compiling and even automatize the procedure to make it less time-consuming. Comparing to other languages even Arabic which is Unicode like Persian, one can say that Persian is left alone in corpus linguistics. It is the time to change that and go further in levels of accuracy to be as close as possible to English NLP tools. Persian NLP needs large amounts of unlimited naturally occurring texts as Persian Today Corpus available in worldwide web to be compiled and accessible to linguists and scientists of related fields.

References

[1] O'Keeffe, A. & McCarthy, M., 2010. *The Routledge Handbook of Corpus Linguistics*. Routledge, London.

[2] Biber, D. & Reppen, R., 2015. *The Cambridge Handbook of English Corpus Linguistics.* Cambridge University Press, Cambridge.

[3] Wilks, Y., 2009. *Machine Translation*. Springer, Sheffield.

[4] Molina, B., 2018. USA Today. [Online]
Available at: https://www.usatoday.com/story/tech/ news/2018/01/16 /robots-better-reading-than-humans/1036420001/ [Accessed 16 January 2018].

[5] M. T. Pilevar, H. Faili, and A. H. Pilevar, "TEP: Tehran English-Persian Parallel Corpus", in Proceedings of 12th *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2011).*

[6] Bijankhan, M., 2004. "The Role of the Corpus in Writing a Grammar: An Introduction to a Software*". Iranian Journal of Linguistics*, *19*(2), pp. 38–67.

[7] Kennedy, H., 2016. *Post, Mine, Repeat Social Media Data Mining Becomes Ordinary*. Macmillan, Sheffield.

[8] Zafarani, R., Abbasi, M. A. & Liu, H., 2014. *Social Media Mining*. Cambridge University Press, New York.
[9] Kerremans, D., 2012. *A Web of New Words*. Peter Lang, Munchen.
[10] Lawson, R., 2015. *Web Scraping with Python*. Packt, Birmingham.
[11] Warden, P., 2011. *Data Source Handbook*. O'Reilly, Cambridge.
[12] Peri, C., 2011. *Teach Yourself the Twitter API*. SAMS, Indianapolis.
[13] Squire, M., 2016. *Mastering Data Mining with Python*. Packt, Birmingham.
[14] Whorf, B. L. & Levinson, J. B., 2012. *Language Thought, and Reality: selected writings of Benjamin Lee Whorf*. MIT Press, Cambridge.
[15] Hald, A., 2007. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*. Springer, New York.
[16] Muller, A. C. & Guido, S., 2017. *Introduction to Machine Learning with Python*. O'Reilly, Boston.
[17] Baker, M., 2000. "Towards a Methodology for Investigating the Style of a Literary Translator". *Target*, 12(2), May, pp. 241-266.
[18] Kao, A. & Poteet, S. R., 2007. *Natural Language Processing and Text Mining*. Springer, London.
[19] Bird, S., Ewan, K. & Edward, L., 2009. *Natural Language Processing with Python*. O'Reilly, Beijing.
[20] Downey, A., 2008. *Think Python*. Green Tea Press, Needham.
[21] Makice, K., 2009. *Twitter API Up and Running*. O'Reilly, Cambridge.
[22] McKinney, W., 2013. *Python for Data Analysis*. O'Reilly, Cambridge.

Appendix

The sample python code which the corpus was compiled with is provided here.

```
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import time
#specified consumer key, consumer secret, access token, access
secret
#which is unique to each app.
ckey="**********"
csecret="**********"
atoken="**********"
asecret="**********"
class listener(StreamListener):
 def on_data(self, data):
 try:
          tweet = data.split(',"text":"')[1].split('","source')[0]
          print (tweet)
          saveThis = str(tweet)
          saveFile = open('corpus.csv','a')
          saveFile.write(saveThis)
          saveFile.write('\n')
          saveFile.close()
          return True
       except BaseException as e:
          print ('failed ondata,',str(e))
          time.sleep(5)
   def on_error(self, status):
       print (status)
auth = OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)
twitterStream = Stream(auth, listener())
```

```
twitterStream.filter(track=["ایران"],languages=['fa'])
```

The python code to encode and decode the compiled corpus is also provided here:

```
sampleFile = open('corpus.csv','r').read()
splitFile = sampleFile.split('\n')
for eachLine in splitFile:
    x = eachLine.encode('utf-8')
    print (x.decode('unicode'))
```

Can Concordle help students to improve reading skills and learning vocabulary?

Azadeh Nemati*

Abstract

This study was based on corpus linguistics. To do this corpus-based study, the first 5 units of the book "concepts and comments" were typed and submitted to concrodle software online program. The output was shown in the form of word clouding and concordancing. The participants were 270 general English students of Islamic Azad University, Jahrom branch, which randomly assigned to control and experimental groups. Word clouding was used in the experimental groups to find main idea and concordacing was used to teach vocabulary as well as reading comprehension skill. While in the control group this intervention was not available and the students did not have access to word clouding and concordancing. Each session, tests of reading comprehension and vocabulary were conducted at the end for both groups. The result of repeated measure ANOVA for five exams revealed that though both groups improved from the first exam to the fifth, this improvement was significant for the experimental group. In other words, control and experimental groups performed differently in each exam in favor of the experimental group. This meant that concordancing had a positive effect on finding main idea, reading skill and vocabulary learning after a repeated time.

Keywords: Concordle, word clouding, reading comprehension, main idea, learning vocabulary

1. Introduction

In recent years a lot of investigations have been devoted to how computers can facilitate language learning [1]. One specific area on computer frontier which still remains open to exploration is corpus linguistics (ibid). In order to conduct a study of language which is

* Department of English language Teaching, Jahrom Branch, Islamic Azad University, Jahrom, Iran, azadehnematiar@yahoo.com

corpus-based, it is necessary to gain access to corpus and concordancing program. A concordance is a software program which analyzes corpora and lists the results. It has a variety of applications such as lexicography and dictionary making. According to Barlow [2], there were three realms in which corpus linguistics can be applied to teaching which were syllabus design, material development and classroom activates. This study will focus on the application of concordancing in language learning.

Another method of teaching in this investigation is by word clouding. A tag cloud (word cloud, or weighted list in visual design) is a visual representation for text data, typically used to depict keyword metadata (tags) on a site [3]. The examples of both concordancing and word clouding were provided in the appendices I and II.

Concorlde is going to be used in this study to provide a corpora to be used in classroom teaching and learning as well as for providing a curriculum. Its effect will be studied on two aspects of language, namely vocabulary and reading comprehension as well as finding the main idea. Vocabulary and comprehension were important in meaningful communication. In fact vocabulary is the best predictor of reading comprehension [4].

Vocabulary has different aspects such as breath, depth and size. It can also be learnt by different methods of intentionally and unintentionally by extensive reading. While student use a lot of time and energy to learn new vocabulary they also forget vocabulary items very soon. In this study by means of concordancing a new method of teaching vocabulary is devised which will be helpful.

Beside vocabulary there is the area of reading comprehension and finding the main idea. Reading is the most frequently mentioned academic subject in which students experience(1982). Reading comprehension on the other hand involves a complicated combination of skills in which students utilize their understanding of various elements, the how of finding main ideas and details and make a distinction between the two [5].

Regarding the importance of the topic of this paper, reading and vocabulary are two important aspects of language. Although different methods were used in the class, students were not still strong in these areas. Using concorancing and word clouding in class will ease teaching of reading comprehendion and vocabulary as it will help students find the main idea of texts. A lot of works have already been done in the realm of vocabulary and reading comprehension but concorlde is a new method that is useful in teaching from which teachers and students could benefit a lot and that is exactly what the author has sought in this paper

1.1 Purposes of the study

There were three main purposes in the present study. First the researcher intended to find the impact of using word clouding in finding the main ideas. Secondly, it intended to find the effect of concordancing in reading comprehension. Finally, it intended to find the effect of concordancing in teaching vocabulary. Based on these purposes the following questions were introduced:

1) What is the effect of word clouding in finding main idea in experimental group?
2) What is the effect of concordance in reading comprehension in experimental group?
3) What is the effect of concordance in teaching vocabulary in experimental group?

And given the above questions, the following hypotheses were formulated:

1) Word clouding has a positive effect in in finding main idea in experimental group.
2) Concordancing has a positive effect in reading comprehension in experimental group.
3) Concordancing has a positive effect in learning vocabulary in experimental group.

2. Review of literature

The role of computer in modern science is well known. Similarly in language study computer analysis of texts reveal some works in this field. In their book under the title of Concordances in the classroom in 1990 Tribbles and Jones mostly talked about the use of concorcancing in teaching.

Stevensens [6] at Sultan Qaboos University, in Oman worked on vocabulary and concordancing and predicted that learners would be able to retrieve a word from memory more successfully by concordance line. Fan [7] investigated the effect of Collaborative Strategic Reading (CSR) on Taiwanese university students' reading comprehension. It was found out that implementing comprehension strategy instruction for one semester may help learners adopt some degree of strategic reading behaviours. Jalalifar [8] investigated the impact of Student Team Achievement Divisions (STAD) and Group Investigation (GI), which were two techniques of Cooperative Learning, on students' reading comprehension achievement of English as a Foreign Language (EFL). The results revealed that STAD is a more effective technique in improving EFL reading comprehension achievement whereas GI and CI did not enhance reading comprehension significantly.

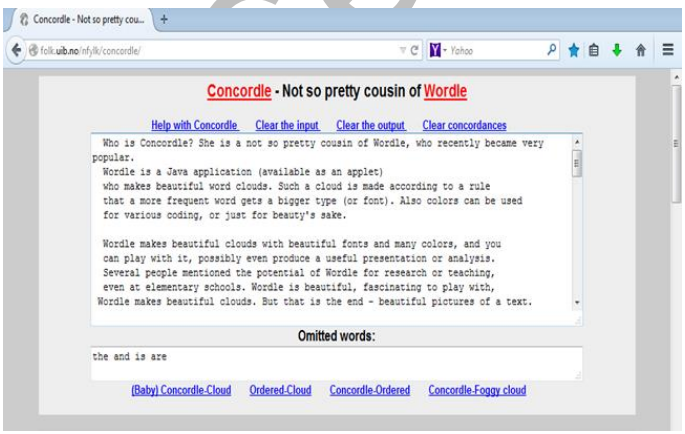
Pakzadian and Esmaili Rasekh [5] explored the effectiveness of using summarization strategies makes any significant difference in EFL learners' level of comprehending English texts. They found that summarization was effective. Wilawn [9] wrote an article about fostering main idea comprehension by cognitive and metacognitive strategies. Rusiecki [10]gave some suggestions for computer assisted teaching of reading in a foreign language. Krieger [1] wrote an article regarding corpus linguistics and explained how it can be applied to teaching. Types, size and other aspect of corpus linguistics was explained by Flowerdew [11].

Though literature is full of related work no one worked on concordancing and reading comprehension.

2.1 Framework of the study

This study is based on corpus linguistics. To do a corpus based study one needs to gain access to corpus and concordancing program. A corpus consists of data banks of natural texts, compiled from writing [1]. And its focus is on authentic language use. To gain this authentic data one way is concordancing. It is a means of accessing a corpus of text to show how any given word or phrase in the text is used in the immediate context in which it appears [11]. There were a lot of software to do concodrancing such as wordle and concorlde .

Concordle, no so pretty cousin of worlde is used in this study to do concordancng. It shows the pattern in which a given word is used. So, learners unconsciously expose to authentic materials. The first five units of the book "concepts and comments" were typed and submitted to concordle, to be used as the corpus of this study. Concordale framework is shown in the next figure and the outputs (word clouding and concrodancing) were shown in Appendices I and II.



3. Methodology

3.1 Participants

The participants of this study were general English students of Islamic Azad University, Jahrom branch. They were selected based on purposive sampling as they had general English course with the researcher. Then they were divided randomly to control and experimental groups. They

were around 270 students from different majors (140 in control and 130 in experimental group) with the mean age of 27.

3.2 Data collection instruments

Two instruments was used in this research as follows:

The first one was concorlde. Who is Concordle? She is a not so pretty cousin of Wordle, who recently became very popular. Wordle is a Java application (available as an applet) who makes beautiful word clouds. Such a cloud is made according to a rule that a more frequent word gets a bigger type (or font). Also colors can be used for various coding, or just for beauty's sake. Wordle makes beautiful clouds with beautiful fonts and many colors, and you can play with it, possibly even produce a useful presentation or analysis. Several people mentioned the potential of Wordle for research or teaching, even at elementary schools. Wordle is beautiful, fascinating to play with, Wordle makes beautiful clouds. But that is the end - beautiful pictures of a text. It can be reached at olk.uib.no/nfylk/concordle/. The concordance at the bottom of the wordle is also used for new items to be shown in concordance form.

The second instrument was the book "concept and comments". This book was purposefully selected because each unit consisted of questions regarding main idea, reading comprehension and vocabulary questions which could be used for both control and experimental groups as test each session.

3.3 Data collection procedure

This study was a true experimental group, as there were both control and experimental groups. The subjects were selected randomly and there were both pre and posttests. In the experimental group word clouding and concordance which were prepared based on their book by concorlde were used. Word clouding helped leaners to get main idea and concordancing was useful for reading skill and vocabulary learning. Then each session after using word clouding and concordancing main idea, reading comprehension as well as vocabulary tests and was used immediately

after teaching. The same tests were used in control group but without the intervention of word clouding and concordancing in class.

4. Results and Discussion

Table 1 presents means and standard deviations of variables in the experimental and control groups.

Table 1. Means and standard deviations of variables in the experimental and control groups

		Experimental Group (N=25)		Control Group (N=25)	
Variables	Repeated Exam	M	SD	M	SD
Main Idea	1	23.56	4.011	22.44	4.610
	2	25.72	4.774	24.56	4.234
	3	34.68	9.168	25.68	4.230
	4	43.56	10.264	31.52	4.063
	5	49.36	4.471	41.76	4.381
Reading Comprehension	1	38.16	8.664	34.84	8.683
	2	39.12	7.886	35.64	6.264
	3	41.20	8.888	36.56	7.985
	4	43.28	8.193	37.32	7.448
	5	43.92	7.984	37.68	7.857
Vocabulary	1	36.28	3.470	36.92	4.830
	2	39.60	4.890	37.32	7.680
	3	44.40	3.175	38.48	7.036
	4	46.80	4.000	39.08	8.195
	5	47.96	2.700	41.40	10.548

The above table showed that in both experimental and control groups the main idea, reading comprehension and vocabulary were increased from the first test to the fifth.

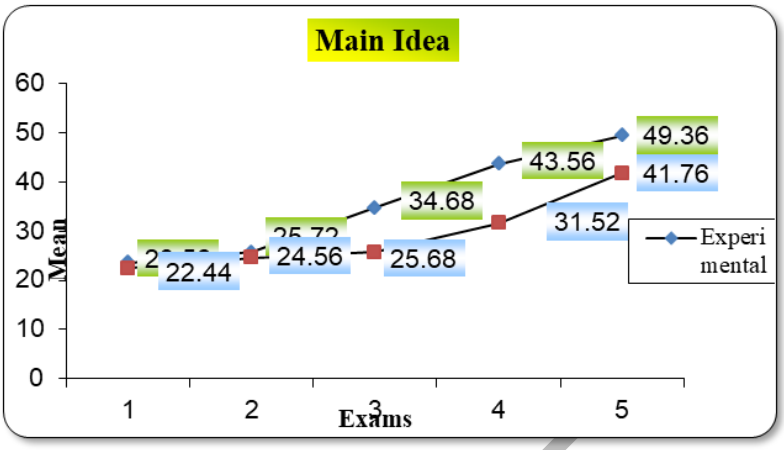


Figure 1: The main idea means in two groups

Figure 1 shows variation of the main idea for experimental and control groups. As seen in the figure, the main idea was increased in both groups and for the experimental group it was rapid after the second exam.

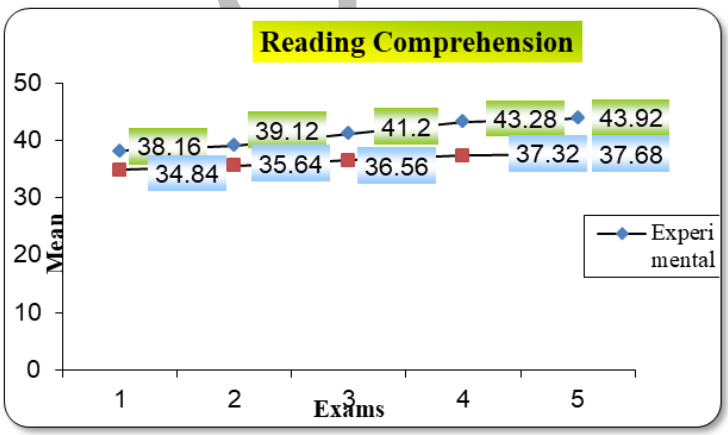


Figure 2: The reading comprehension means in two groups

Figure 2 illustrated variation of the reading comprehension for experimental and control groups. As seen in the figure, the main idea

was slowly increased in both groups and the increasing in the experimental group was higher than the control group.

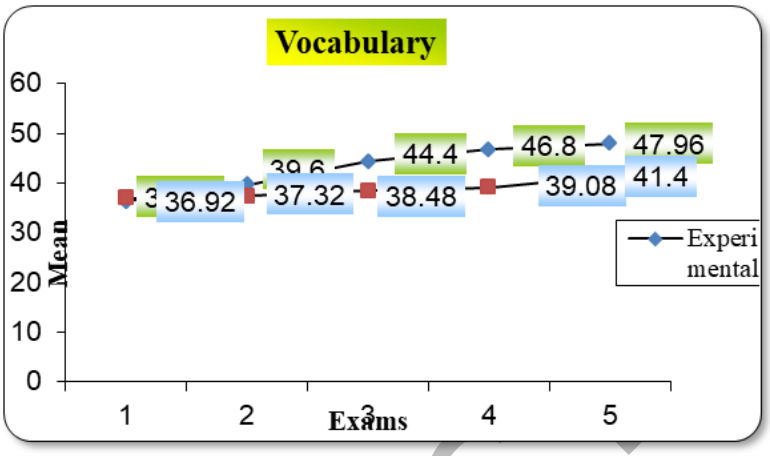


Figure 3: The vocabulary means in two groups

Figure 3 showed that the vocabulary means for both groups were increasing. The increasing in the experimental group was higher than the control group. Prior to investigate research questions, the normality of variables by the Kolmogorov-Smirnov test was examined.

Table 2. Kolmogorov-Smirnov test of normality in the experimental and control groups

		Experimental Group (N=25)		Control Group (N=25)	
Variables	Repeated Exam	K-S	Sig.	K-S	Sig.
Main Idea	1	1.057	0.214	0.561	0.911
	2	0.665	0.769	0.791	0.558
	3	0.632	0.819	0.607	0.855
	4	0.781	0.575	0.797	0.550
	5	0.710	0.695	0.601	0.863
Reading Comprehension	1	0.537	0.935	0.759	0.612
	2	1.130	0.156	0.618	0.840

		Experimental Group (N=25)		Control Group (N=25)	
	3	0.761	0.609	0.761	0.609
	4	0.676	0.751	0.550	0.923
	5	0.754	0.620	0.665	0.769
Vocabulary	1	0.542	0.930	0.581	0.889
	2	0.583	0.886	0.863	0.445
	3	0.518	0.951	0.675	0.753
	4	0.854	0.460	0.800	0.545
	5	0.921	0.364	0.605	0.857

Table 2 presents result of the Kolmogorov-Smirnov test and indicated that the test results were not significant (Sig.>0.05) for all variables in both groups and in any stage of the exam. Non-significant results meant that the variables distribution were normal.

Question 1: What is the effect of word clouding in finding main idea in experimental group?

To answer this question the repeated measure analysis was applied. The normality condition of the variable distribution was satisfied.

Table 3. Repeated measure analysis for the main idea variable

Source	Wilks' Lambda	F	df1	df2	p	η^2
Repeated factor	0.037	292.906	4	45	0.001	0.963
Repeated factor * group	0.401	16.835	4	45	0.001	0.599

The repeated measure analysis was done by the Wilks' Lambda statistics and results indicated that the effect of repeated factor was significant (Sig.<0.05). Thus there was a significant difference in the main idea from the first test to the fifth. Also the interaction of the repeated factor and group was significant (Sig.<0.05) which indicated that two groups were differently performed during repeating the exams of the main idea.

The eta squared equaled to 0.963 and showed a huge effect size of the repeated factor on the main idea. In other words, the significant difference of the main idea in repeated exams was very considerable. The effect size of the interaction between repeated factor and group was η2=0.599 which is a big effect size.

Table 4. Means and confidence intervals for the main idea variable in two groups

Group	Mean	Std. Error	95% Confidence Interval		Pairwise Comparisons	
			Lower Bound	Upper Bound	Mean Difference	Sig.
Control	29.192	0.931	27.320	31.064	6.184	0.001
Experimental	35.376	0.931	33.504	37.248		

Table 4 presented the overall means of the main idea in experimental and control groups. It showed that the experimental group's mean was higher than the confidence interval of the control group and the mean difference was significant (Sig.<0.05).

Table 5. Means and confidence intervals for the main idea variable in repeated factor and groups

Group	Repeated exam	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Control	1	22.440	0.864	20.702	24.178
	2	24.560	0.902	22.746	26.374
	3	25.680	1.428	22.809	28.551
	4	31.520	1.561	28.381	34.659
	5	41.760	0.885	39.980	43.540
Experimental	1	23.560	0.864	21.822	25.298
	2	25.720	0.902	23.906	27.534
	3	34.680	1.428	31.809	37.551

Group	Repeated exam	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
	4	43.560	1.561	40.421	46.699
	5	49.360	0.885	47.580	51.140

As seen in Table 5, the main idea scores, for both groups, were increased in each of the repeated exam and means at each exam was higher than the previous exam confidence interval. This meant that the increasing of the main idea was significant between two successive exams.

Question 2: What is the effect of concordance in reading comprehension in experimental group?

To answer this question the repeated measure analysis was used. The normality condition of the variable distribution was satisfied.

Table 6. Repeated measure analysis for the reading comprehension variable

Source	Wilks' Lambda	F	df1	df2	p	η^2
Repeated factor	0.307	25.449	4	45	0.001	0.693
Repeated factor * group	0.740	3.961	4	45	0.008	0.260

The repeated measure analysis was done by the Wilks' Lambda statistics and results indicated that the effect of repeated factor was significant (Sig.<0.05). Thus there was a significant difference in the reading comprehension from the first test to the fifth. Also the interaction of the repeated factor and group was significant (Sig.<0.05) which indicated that two groups were differently performed during repeating the exams of the Reading Comprehension.The eta squared equaled to 0.693 and showed a big effect size of the repeated factor on the reading comprehension. In

other words, the significant difference of the reading comprehension in repeated exams was very considerable. The effect size of the interaction between repeated factor and group was η2=0.260 which is a moderate effect size.

Table 7. Means and confidence intervals for the reading comprehension variable in two groups

Group	Mean	Std. Error	95% Confidence Interval		Pairwise Comparisons	
			Lower Bound	Upper Bound	Mean Difference	Sig.
Control	36.408	1.522	33.349	39.467	4.728	0.033
Experimental	41.136	1.522	38.077	44.195		

Table 7 presented the overall means of the reading comprehension in experimental and control groups. It showed that the experimental group's mean was higher than the confidence interval of the control group and the mean difference was significant (Sig.<0.05).

Table 8. Means and confidence intervals for the reading comprehension variable in repeated factor and groups

Group	Repeated exam	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Control	1	34.840	1.735	31.352	38.328
	2	35.640	1.424	32.776	38.504
	3	36.560	1.690	33.163	39.957
	4	37.320	1.566	34.172	40.468
	5	37.680	1.584	34.495	40.865
Experimental	1	38.160	1.735	34.672	41.648
	2	39.120	1.424	36.256	41.984
	3	41.200	1.690	37.803	44.597

Group	Repeated exam	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
	4	43.280	1.566	40.132	46.428
	5	43.920	1.584	40.735	47.105

As seen in Table 8, the reading comprehension scores, for both groups, were increased in each of the repeated exam. Means at each exam was slightly higher than the previous exam. The overall increasing from the first exam to the fifth was significant and the fifth reading comprehension mean was higher than the first confidence interval for each of groups.

Question 3: What is the effect of concordance in teaching vocabulary in experimental group?

To answer this question we used the repeated measure analysis. The normality condition of the variable distribution was satisfied.

Table 9. Repeated measure analysis for the vocabulary variable

Source	Wilks' Lambda	F	df1	df2	p	η^2
Repeated factor	0.365	19.573	4	45	0.001	0.635
Repeated factor * group	0.628	6.665	4	45	0.001	0.372

The repeated measure analysis was done by the Wilks' Lambda statistics and results indicated that the effect of repeated factor was significant (Sig.<0.05). Thus there was a significant difference in the Vocabulary from the first test to the fifth. Also the interaction of the repeated factor and group was significant (Sig.<0.05) which indicated that two groups were differently performed during repeating the exams of the Vocabulary. The eta squared equaled to 0.635 and

showed a big effect size of the repeated factor on the vocabulary. In other words, the significant difference of the vocabulary in repeated exams was very considerable. The effect size of the interaction between repeated factor and group was η2=0.372 which is a moderate effect size.

Table 10. Means and confidence intervals for the vocabulary variable in two groups

Group	Mean	Std. Error	95% Confidence Interval		Pairwise Comparisons	
			Lower Bound	Upper Bound	Mean Difference	Sig.
Control	38.640	1.043	36.544	40.736	4.368	0.005
Experimental	43.008	1.043	40.912	45.104		

Table 10 presented the overall means of the Vocabulary in experimental and control groups. It showed that the experimental group's mean was higher than the confidence interval of the control group and the mean difference was significant (Sig.<0.05).

Table 11. Means and confidence intervals for the vocabulary variable in repeated factor and groups

Group	Repeated exam	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Control	1	36.920	0.841	35.229	38.611
	2	37.320	1.288	34.731	39.909
	3	38.480	1.092	36.285	40.675
	4	39.080	1.290	36.487	41.673
	5	41.400	1.540	38.304	44.496
Experimental	1	36.280	0.841	34.589	37.971
	2	39.600	1.288	37.011	42.189

3	44.400	1.092	42.205	46.595
4	46.800	1.290	44.207	49.393
5	47.960	1.540	44.864	51.056

As seen in Table 11, the vocabulary scores, for both groups, were increased in each of the repeated exam. Means at each exam was slightly higher than the previous exam. The overall increasing from the first exam to the fifth was significant and the fifth vocabulary mean was higher than the first confidence interval for each of groups.

5. Conclusions

The role of computer in different sciences such as biology, physics, chemistry etc. is well known. Computer has also affected language teaching and learning. The computational analysis of language began in the 1960. However, one new aspect of computational analysis is corpus linguistics such as concodle and wordle. Concordancing has a variety of applications. One earliest of these was in the field of lexicography and dictionary learning [11]. The result of which is Collins cobuild dictionary.

In this study three research questions were proposed to be answered regarding the application of word clouding and concordancing to improve reading skills and vocabulary learning. The results indicated that concordancing was effective to find main idea better, improving reading skills and learning vocabulary more. As Rusiecki [10]stated that “concordancing can be used as the foundation for a program for teaching reading”. Then he added that context in which a new word is found does not necessarily contain sufficient information, while concordance output by presenting several context of the same word simultaneously greatly increase the chance of success. The results of the preset study indicated that by grouping the uses of particular word the concordance showed the pattern in which a given word is used. As a result it is a good hand for learners to specify for them what they

should look at and why they should look at it. It is an effective way of encoding for them the unknown part of language.

References

[23] Krieger, D. (2013). Corpus lingusitics: What is and how it can be applied to teaching. Retreived 23 May, from iteslj.org/Articles/ Krieger-Corpus.htm.

[24] Barlow, M. (2002). Corpora, concordancing and language teaching. *Proceedings of the 2002 KAMALL international conference*. Daejon, Korea.

[25] Wikepedia, (2014). Tag cloud. Retrieved 8 May, from en.wikipedia.org/wiki/Tag_cloud.

[26] Nation, I. S. P (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

[27] Pakzadian, M., & Eslami Rasekh, A. (2012). The effects of using summarization strategies on Iranian EFL learners' reading comprehension. *English Linguistics Research*, *1*(1), 118-125.

[28] Stevensens, V. (1991). Concordance-based vocabulary exercises: A viable alternative to gap-fillers. In Johns, T. & King, P. (eds.) *Classroom concordancing: English Language Research Journal*, 4 (47-63). University of Birmingham: Centre for English Language Studies.

[29] Fan, Y. (2010). The Effect of comprehension strategy instruction on EFL learners' reading comprehension, *Asian Social Science*, *6*(8), 19-29.

[30] Jalalifar, A. (2012). The effect of cooperative learning techniques on college students' reading comprehension. *System 38*, 96-108.

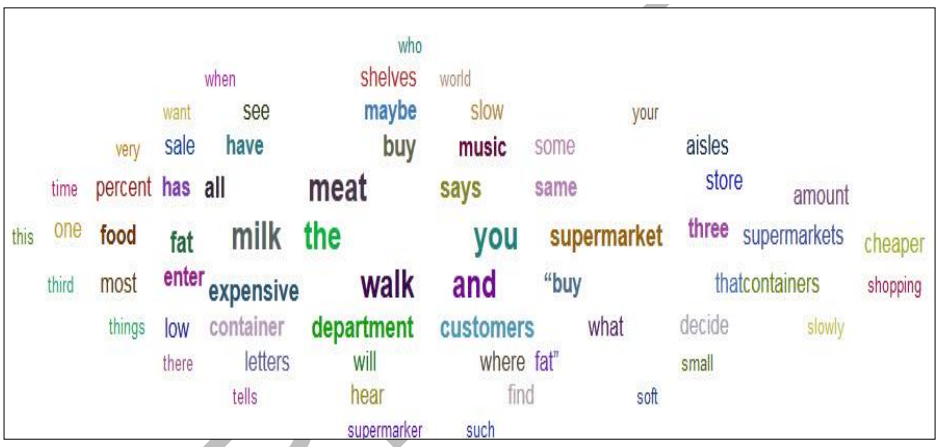
[31] Wilawan, S. (2012). Fostering main idea comprehension among EFL learners through cognitive and metacognitive strategies. *International Journal of Humanities and Social sciences*, *2*(14), 46-54.

[10]. Rusiecki, J. (2002). Context, concordance, and what next? Suggestions for computer-assisted teaching of reading in a

foreign language. *Teaching English with Technology*, *2*(2), 5-12. Also available from http://tewtjournal.webs.com/VOL%202/ISSUE%202/VOL%202%20ISSUE%202%20COMPLETE.pdf.

[11]. Flowerdew, J. (2014). Concordancing in language learning. Retrieved 10 June, from citeseerx.ist.psu.edu/viewdoc/download? doi=10.1.1.127.3263...

Appendix I: word clouding for unit one of the book concept and comments



Appendix II: concordancing for different words of unit one of the book concept and comments

you buy in the supermarker? Do you	decide?	Does the supermarket decide? When you enter
supermarker? Do you decide? Does the supermarket	decide?	When you enter the supermarket, you see
it. One store has three differ - ent	containers	of low fat milk. One says "1

same. However, in this store the three	containers	of milk cost Three different amounts of
cities all over the world shop in	supermarkets.	Who decides what you buy in the
the most. Most of the food in	supermarkets	is very attractive. It all says “Buy
see, all the milk has the same	amount	of fat. The milk is all the
The milk is all the same. The	amount	of milk in each container is also
full of food. You walk in the	aisles	between the shelves. You push a shopping
slow music as you walk along the	aisles.	If you hear fast music, you walk
only a little butterfat in it. One	store	has three differ -ent containers of low
is also the same. However, in this	store	the three containers of milk cost Three
little butterfat in it. One store has	three	differ -ent containers of low fat milk.
the same. However, in this store the	three	containers of milk cost Three different amounts
store the three containers of milk cost	Three	different amounts of money. Maybe the customer
the world shop in supermarkets. Who decides	what	you buy in the supermarker? Do you
less fat. ” The supermarket tells you	what	to buy.
supermarkets is very attractive. It all says	“Buy	me! ” to the customers. The expensive
to the customers. The expensive meat says	“Buy	me! ” as you walk by. The
walk by. The expensive milk	“Buy	me! I have less fat. ” The

container says		
can see, all the milk has the	same	amount of fat. The milk is all
of fat. The milk is all the	same.	The amount of milk in each container
milk in each container is also the	same.	However, in this store the three containers
the supermarker? Do you decide? Does the	supermarket	decide? When you enter the supermarket, you
the supermarket decide? When you enter the	supermarket,	you see shelves full of food. You
hear fast music, you walk quickly. The	supermarket	plays slow music. You walk slowly and
to find it. The manager of the	supermarket	knows where customers enter the meat department.
me! I have less fat. ” The	supermarket	tells you what to buy.
in it. You probably hear soft, slow	music	as you walk along the aisles. If
along the aisles. If you hear fast	music,	you walk quickly. The supermarket plays slow
you walk quickly. The supermarket plays slow	music.	You walk slowly and have more time
-ent containers of low fat milk. One	says	“1 percent (1 %) fat” on the
%) fat” on the container. The second	says	“ 99 percent (99 %) fat free.
in supermarkets is very attractive. It all	says	“Buy me! ” to the customers. The
” to the customers. The expensive meat	says	“Buy me! ” as you walk by.
you walk by. The expensive milk container	says	“Buy me! I have less fat. ”

The manager of the supermarket knows where	customers	enter the meat department. The cheaper meat
the meat department, away from where the	customers	enter. You have to walk by all
products such as butter and cheese. Many	customers	like milk that has only a little
all says “Buy me! ” to the	customers.	The expensive meat says “Buy me! ”
it. The manager of the supermarket knows	where	customers enter the meat department. The cheaper
end of the meat department, away from	where	the customers enter. You have to walk
When you enter the supermarket, you see	shelves	full of food. You walk in the
You walk in the aisles between the	shelves.	You push a shopping cart and put
and have more time to buy things.	Maybe	you go to the meat department first.
meat before you find the cheaper meat.	Maybe	you will buy some of the expensive
milk cost Three different amounts of money.	Maybe	the customer will buy the milk that
shop in supermarkets. Who decides what you	buy	in the supermarker? Do you decide? Does
walk slowly and have more time to	buy	things. Maybe you go to the meat
find the cheaper meat. Maybe you will	buy	some of the expensive meat instead of
amounts of money. Maybe the customer will	buy	the milk that costs the most. Most
” The supermarket tells you what to	buy.	

you find the cheaper meat. Maybe you	will	buy some of the expensive meat instead
different amounts of money. Maybe the customer	will	buy the milk that costs the most.
put your food in it. You probably	hear	soft, slow music as you walk along
you walk along the aisles. If you	hear	fast music, you walk quickly. The supermarket
third saya “ Low Fat” in bing	letters	and “1 % ” in small letters.
letters and “1 % ” in small	letters.	As you can see, all the milk
plays slow music. You walk slowly and	have	more time to buy things. Maybe you