پارسی ست: تولید مجموعه‌ی ابزار تجزیه کننده‌ی جمله به واژه و پیکره‌ی مجموعه واژگان مبهم برای زبان فارسی

دکتر محمدباقر دستغیب

گروه پژوهشی طراحی و عملیات سیستم‌ها

آذرماه ۱۳۹۷

بسم الله الرحمن الرحیم

فهرست مطالب

فهرست جداول

عنوان جدول **صفحه**

فهرست تصاویر

عنوان **صفحه**

چکیده

زبان فارسی یکی از زبان‌های دنیا است که نظام نوشتاری آن الفبایی است که از زبان عربی برگرفته شده است. حدود یک درصد جمعیت جهان فارسی زبان هستند و همین امر، انجام تحقیقات زبان‌شناسی در این حوزه در سطح ملی و فرا ملی را بسیار با اهمیت می‌سازد. انجام پژوهش در حوزه‌ی زبان‌شناسی رایانشی در حوزه‌ی هر زبان، نیازمند ابزار و منابع زبانی (پیکره‌ها) است. بنابراین تهیه ابزارها و پیکره‌ها از جمله پیش‌نیازهای پژوهش در حوزه‌ی زبان‌شناسی رایانه‌ای است. در این راستا در این پژوهش مجموعه‌ی پارسی‌ست که شامل تجزیه کننده‌ی جمله به واژه و پیکره‌ی مجموعه‌ی ابهام واژگان فارسی است، تولید شده است. پیکره‌ی مجموعه‌ی ابهام برای هر واژه‌ی صحیح فارسی، شامل مجموعه‌ی واژگان صحیح فارسی در فاصله‌ی ویرایشی یک نسبت به آن واژه است. این مجموعه‌ی ارزشمند می‌تواند در شناسایی نوری نویسه‌ها، تصحیح وغلط‌یابی متون فارسی و تبدیل گفتار به متن مورد استفاده قرار گیرد. محصول دیگر پارسی ست، تجزیه کننده است که وظیفه‌ی آن تجزیه‌ی جمله به واژه به روش ترکیبی است. این تجزیه کننده همچنین با کمک روش مبتنی بر دانش، نیم‌فاصله را نیز اصلاح می‌نماید.

کلیدواژه: پارسی‌ست، تجزیه‌کننده متن، روش مبتنی بر دانش، پیکره‌ی مجموعه‌ی ابهام

فصل اول

مقدمه

۱-مقدمه

زبان فارسی زبان رسمی ایران دارای نظام نوشتاری الفبایی است که برگرفته از زبان عربی، بعلاوه‌ی چهار حرف (پ،ژ،گ و چ) است. نوشتار به زبان فارسی از راست بچپ نوشته می‌شود و در دنیا بیش از ۱٪ جمعیت جهان فارسی را به عنوان زبان اول یا دوم بکار می‌برند. با پیشرفت تکنولوژی و بکارگیری دستگاه‌های الکترونیکی مانند رایانه‌ها، گوشی‌های هوشمند و ... تولید متون رقمی در زبان‌های مختلف با سرعت بسیاری در حال انجام است. حال آنکه بکارگیری و استفاده از داده‌های ارزشمندی که بصورت رقمی تولید می‌شود، می‌تواند خدمات ارزشمندی به سهولت برای کاربران را فراهم نماید.

تولید الکترونیکی متون در تمامی حوزه‌ها از جمله سطح اینترنت موجب افزایش نیاز به ابزارهای زبان‌شناسی شده است. بنابراین باید برای ساخت ابزارهای زبان‌شناسی در حوزه‌ی فارسی برنامه‌ریزی شود. در تمامی زبان‌های رایج دنیا، ابزارهای پردازش زبان به عنوان ابزارکهای کاربردی رسماً منتشر شده‌اند. برای زبان فارسی دسترسی به منابع و ابزارهای زبان از جمله چالش‌های پیش روی پژوهشگران در حوزه‌ی زبان شناسی است.

زبان فارسی، به سبب ویژگیهای خاص آن و در عین حال نهادینه نشدن سبک نگارش استاندارد، در رویارویی با محیطهای الکترونیکی، با دشواریهایی روبه‌روست که تأثیری بسزا بر اثربخشی بازیابی اطلاعات می‌گذارد. پژوهش حاضر می‌کوشد تا با بررسی متون و پیشینه‌های موجود، چالشهای نگارش فارسی، تأثیر آنها بر اثربخشی بازیابی اطلاعات، و پیشنهادهای ارائه شده در جهت رفع این دشواریها را مورد بحث و بررسی قرار دهد. با تحلیل و مرور جامع متونی که دربارهٔ چالشهای نگارش فارسی در محیطهای دیجیتال نگارش یافته است، می‌توان دانشی

را که تاکنون در این باره گرد آمده است به تصویر کشید و کاستی‌ها و پیشرفت‌های به دست آمده در این زمینه را آشکار ساخت.

خواندن و نگارش فارسی به دلیل ویژگیهای خاصّ این زبان، در پاره‌ای موارد با دشواریهایی همراه است که در رویارویی با رایانه، دو چندان می‌گردد. ورود ناگهانی رایانه به گستره‌ای وسیع از فعالیتهای مختلف اجتماعی، فرهنگی، اقتصادی و فنی، مجال آن را به صاحب نظران نداده است که راهکاری بنیانی و جامع برای مقابله با چالشهای شیوهٔ نگارش بیندیشند و به کار گیرند (ستوده، ۱۳۹۱). نبود استاندارد شیوهٔ نگارش جامع و مورد قبول همگان، به نایکدستی و ناهماهنگی دروندهی اطلاعات در پایگاه‌های اطلاعاتی، وب‌سایت‌ها، وب‌لاگ‌ها و دیگر منابع دیجیتالی انجامیده که آن نیز به نوبهٔ خود جستجوی فارسی را با مشکلاتی چند همراه ساخته است. این دشواریها بویژه در دنیای وب و با رشد سریع انتشارات الکترونیکی فارسی بر وب، چشمگیر بوده است. شیوه‌نامه‌ای که فرهنگستان ادب و زبان فارسی در سالهای اخیر برای یکدستی نگارش فارسی ارائه کرده نیز نتوانسته است از این دشواریها بکاهد، زیرا این شیوه‌نامه به دلیل ناهماهنگی درونی، هدف قرار دادن عامه مردم و در نتیجه کاهش دقت و پرهیز از وضع قانون برای برخی استثناها، وضع قانون برای پیوسته یا جدانویسی برخی کلمات مرکب و واگذار کردن سایر موارد به سلیقهٔ نویسندگان و در نهایت نپرداختن به همهٔ دشواریهای نگارشی، مورد انتقاد بوده است. از سوی دیگر، عدم بازخواست به کارگیری این دستورها باعث می‌شود پذیرش و نهادینه شدن این سبک، فرایندی بسیار بلندمدت، اگر نگوییم ناشدنی، باشد (ستوده، ۱۳۹۱).

دسترسی آسان به انبوهی از اطلاعات، دستاورد حضور اطلاعات در محیطهای الکترونیکی بخصوص وب است. در کنار این مزیّت، مسئله بازیابی اثربخش اطلاعات رخ می‌نماید. اثربخشی بازیابی زمانی حاصل می‌شود که نیاز کاربر هرچه بیشتر و بهتر برآورده گردد؛ بدین معنا که شمار بیشتری از مدارک با درجه ربطِ هرچه بیشتر با

موضوعِ مورد نظر وی بازیابی گردد. اهمیت این مسئله زمانی که اطلاعات به زبانی چون فارسی مورد نیاز باشد، دوچندان می‌گردد. زیرا شیوهٔ نگارش زبان فارسی، به سبب ویژگیهای خاصّ آن و در عین حال نداشتن سبکی استاندارد، در رویارویی با محیطهای الکترونیکی، با دشواریهایی روبه‌روست که تأثیری بسزا بر اثربخشی بازیابی اطلاعات می‌گذارد (ستوده، ۱۳۹۱).

وجود ارتباط متقابل میان زبان‌شناسان و فنآوران حوزه ی رایانه، یکی از نیازهای اصلی جامعه‌ی اطلاعاتی امروز و رشد صنعت پردازش الکترونیکی در کشور ایران است (Kashefi, Nasri, & Kanani, 2010). اگر در ساختار خط فارسی تغییری برای همگام شدن با این رشد روی ندهد، میزان عقب ماندگی کشور ایران در فنآوری اطلاعات از سایر کشورها جبران ناپذیر خواهد شد. زمانی فرا خواهد رسید که روزنامه‌ها و مقالات خارجی با سرعت بالایی تولید شده و خلاصه‌ی آن‌ها در کسری از ثانیه استخراج می‌شود، در حالی که ما در ایران حتی کار خطایابی واژگانی را نیز با خطای بالا و به کندی انجام می‌دهیم. زمانی که موتورهای جستجوی دیگر زبان‌ها می‌توانند هنگام جستجوی یک واژه، مترادف‌ها، ریشه‌ها و سایر مشتقات آن را برای زبان‌های غیرفارسی نیز بازیابی کنند، ما همچنان در حال رفع مشکل چندگانگی کدکاراکتر «ی» و یا مشکل اجزای واژه هستیم و یک جستجوی ساده را نیز نمی‌توانیم در وبگاه‌های رسمی کشور به درستی انجام دهیم. از این رو اهمیت موضوع یکسان‌سازی نحوه‌ی نگارش واژه‌ها و نیز موارد ابهام‌زای موجود در دستور خط فارسی دوچندان می‌شود و باید تصمیمی جدی جهت رفع اشکالات و چالش‌های این بخش اندیشید(Kashefi et al., 2010).

بنابراین یکی از چالش‌های تحقیق در حوزه‌ی پردازش زبان طبیعی کمبود منابع و ابزارها است. زیرا برای زبان فارسی ابزارهای پردازش زبان طبیعی در دسترس نیست و خود محقق می باید این مهم را تهیه نماید. این

ابزارها پیشنیاز شروع تحقیق در حوزه‌ی پردازش زبان طبیعی است. ساخت پیکره‌ها برای زبان فارسی نیز کاربردهای فراوان در پردازش زبان طبیعی دارد و یکی از کاربردهای آن، ساخت تجزیه کننده متن است که براساس مدل آماری و قوانین زبان کار می‌کند.

نیاز به وجود ابزارها و به خصوص پیکره‌ها، برای زبان فارسی، لزوم انجام تحقیق و پژوهش در این حوزه را نشان می‌دهد. تلاش‌هایی کم و بیش در این حوزه انجام شده است، ولی معمولاً یا بصورت عمومی در دسترس محققین و علاقه مندان نیست و یا استفاده از آن مستلزم کسب اجازه و طی طریق از طرف مولف است. هدف این تحقیق این است که ابزارها و پیکره حاصل از پژوهش به صورت رایگان در اختیار محققین و علاقه مندان قرار گیرد تا بتوان در این حوزه علاوه بر بهبود کیفیت پژوهش حاضر، به ادامه‌ی تحقیق و پژوهش در زمینه‌ی پردازش زبان طبیعی برای زبان فارسی‌یاری رساند(Kashefi et al., 2010). با توجه به چالش‌های پردازش زبان طبیعی در فارسی برخی از این چالش‌ها مستقلاً مورد بررسی قرار می‌گیرد.

۱-۲-چاش‌های پردازش زبان فارسی

متون نوشته شده به فارسی راست به چپ است و از نظام نوشتاری با الفبایی برگرفته از زبان عربی استفاده می‌کند. یکی از چالش‌های زبان فارسی، واژه شناسی غنی و پیچیده آن است (Kashefi et al., 2010). واژه‌های زبان فارسی با ترکیب‌های بسیاری از پسوندها و پیشوندها صرف می‌شوند. کلمات اشتقاقی بسیاری در زبان فارسی رایج است و به کار برده می‌شود، ولی قوانین اشتقاق، تصریف، و ترکیب آنها دقیق و جامع نیست. ترکیب وندها با اسامی در زبان فارسی یکی از مشکلات شناخت واژه‌ها است ولی بطور بایسته به آن پرداخته نشده و قوانین جامعی برای آن تدوین نشـــده است (Kashefi et al., 2010)

افعال فارسی قوانین پیچیده‌ای برای صرف و ترکیب دارد. علاوه بر آن، وجود افعال مرکب، عدم تولید قوانین رایانه‌ای برای صرف افعال، فعل‌های پیچیده‌ی چند جزئی و همچنین وجود استثناهای فراوان از جمله چالش‌های فعل فارسی است (Megerdoomian, 2000).

زبان فارسی علاوه بر فاصله گذاری معمول در دیگر زبان‌ها که به عنوان جداساز کلمات استفاده می‌شود، فاصله‌ی دیگری بنام نیم‌فاصله و یا شبه فاصله[1] دارد. شبه فاصله‌ها دارای قوانین مدون و دقیقی نیستند.

با توجه به تمامی موارد فوق، تشخیص واژه‌های یک ترتیب یا جمله خود یکی از مهمترین چالش‌های زبان فارسی است. برای زبان فارسی ابزار دقیق و مستندی برای تجزیه جمله به کلمه وجود ندارد و یا به طور عمومی در دسترس محققین نیست و یا در صورت انتشار مشکلات زیادی دارد. این ابزار مورد نیاز بسیاری از پژوهش‌ها در حیطه‌ی پردازش‌های زبان طبیعی است و یک پیشنیاز تحقیق در حوزه پردازش زبان طبیعی محسوب می‌شود. البته در این پژوهش چالش‌های زبان فارسی از دید نوشتاری املاء لغات، و همچنین پردازش رایانه‌ای مورد بررسی قرار گرفته است.

پردازش واژگانی کلیه‌ی زبان‌های طبیعی امری دشوار است. ترکیب واژگان، منجر به تشکیل واژگانی می‌شود که ممکن است در اثر بی‌دقتی کاربران، از دید رایانه به دو یا چند شکل مختلف خوانده شوند. مثلاً در جایی که منظور نویسنده «سیستم‌عامل» است، اگردر اثر بی‌دقتی «سیستم عامل» نوشته شود، رایانه قادر به تشخیص واژه‌ی اصلی نخواهد بود. دومین دلیل پیچیدگی پردازش واژگانی زبان‌های طبیعی، ترکیب واژه‌ها با هم و تولید واژه‌هایی است که حاوی اطلاعاتی مانند مالکیت، جمع یا مفرد بودن واژه (به عنوان نمونه «کتاب‌هایشان») هستند. این واژه‌های جدید در واژه‌نامه‌ها وجود ندارد، اما معنای آن‌ها همان معنایی است، که

[1] Pseudo space

در واژه ی اولیه نهفته بوده است (Anvari & Givi, 2006). از دید رایانه تنها در صورتی دو واژه با هم یکسان هستند که به یک صورت نوشته شده باشند. سومین دلیل مشکل بودن تفسیر واژه از دید رایانه، آن است که برخی از قاعده‌های تولید واژه در زبان‌های طبیعی، می‌توانند واژه‌هایی به وجود آورند که در واژه‌نامه ها وجود ندارند (Shamsfard, Jafari, & Ilbeygi, 2010). از سوی دیگر، اگر رایانه بتواند تمام قاعده‌های ساخت واژه‌ها را در خود جای دهد، در آن صورت واژگانی که امکان تولید آن‌ها در زبان وجود دارد، اما درمتن کاربردی ندارد نیز در فهرست واژه‌های مورد تایید رایانه قرار خواهند گرفت. بنابراین، پردازش واژه‌های یک متن به خودی خود رایانه را با مشکلاتی در تشخیص واژه‌ها مواجه می‌کند. این موارد با صرف نظر از اشتباهات دستور خط فارسی و املایی کاربران در نظر گرفته شده است و فرض شده است تمام کاربران قواعد و اصول نسبتاً یکسانی را در نگارش خود به کار برند (Kashefi et al., 2010). ولی مسلماً با توجه به طیف گسترده‌ی کاربران، این امری محال و نشدنی است و باید در زمان تولید متن استانداردهای تولید رایانه‌ای متن در نظر گرفته شود.

در زبان فارسی، مصوت‌ها با وجود آن که تلفظ می‌شود، اکثراً نوشته نمی‌شود. اگر واژه‌ای با اعراب نوشته شود، از دید رایانه با حالت بی‌اعراب آن متفاوت خواهد بود (Karimi, Tabrizi, & Chalak, 2016). در نتیجه، لازم است که همواره برای واژه‌های با اعراب روش جداگانه‌ای در نظر گرفته شود. از طرفی، اکثر فارسی‌نویسان از گذاشتن برخی از حرکت‌های الزامی مانند تشدیدهای لازم و تنوین خودداری می‌کنند. نتیجه‌ی این امر، متفاوت بودن صورت ظاهری واژه در رایانه، با آنچه در واژه‌نامه‌ها وجود دارد، است. به دلیل آموزش‌های متفاوتی که از طریق نظام آموزش کشور به فارسی‌نویسان داده شده است، حفظ یکپارچگی خط فارسی برای نویسندگان مختلف که آموزش‌های متفاوتی دیده‌اند، مشکل است (Karimi et al,

2016). در زبان‌های دیگر تغییر دستور خط بسیار به ندرت رخ می‌دهد. در نتیجه، تمام مردم در سنین مختلف از طرز صحیح نگارش واژه‌ها و دستور خط خود، آگاهی کامل دارند. البته آگاهی داشتن دلیل بر رعایت کردن اصول نیست، اما وجود توافق کلی در نگارش واژه‌ها امری است که موجب تصحیح متون رسمی و یکدستی خط می‌شود.

ویراستاران حرفه‌ای نیز با وجود آشنایی با دستور خط فارسی و اصول ویرایش، توافق کلی در نگارش واژه ها ندارند. مثلاً ممکن است متن نهایی ویراستاری شده در رایانه توسط چندین مصحح دارای یک توکن به صور گوناگون باشد و این مشکل بزرگی در پردازش‌های رایانه‌ای است (Mahboubi, Compton, & LU, 2017). این کار باعث می‌شود تا نتیجه ی تحلیل متنی که‌یک ویراستار حرفه‌ای به رایانه داده است، مانند نتیجه‌ی تحلیل متن مشابه‌ای که ویراستار حرفه‌ای دیگری آن را ویرایش کرده است، نباشد. این دوگانگی می‌تواند مشکلات زیادی را برای موتورهای جستجو و سیستم‌های تحلیل متن پدید آورد، زیرا یک واژه به صورت‌های متنوع نوشته می‌شود.

مهمترین چالش در حوزه متون رقمی، مسئله‌ی فاصله گذاری است (Shamsfard et al., 2010). تا زمانی که ابهام این این بخش به طور دقیق برای کاربران رایانه حل نشود، متون تولیدی این کاربران برای مراحل بعدی پردازش در رایانه مناسب نخواهد بود و در نتیجه تنها کاربردشان این است که توسط کاربر دیگری خوانده شوند. در این حالت، جستجو، محاسبه‌ی میزان پیچیدگی، خلاصه‌سازی خودکار، یا تشخیص و تصحیح خودکار خطاهای املایی در این متون، مشکلات بسیاری را دربر خواهد داشت (Kashefi et al., 2010). تا زمانی که دو کاربر ورزیده و دانش آموخته‌ی زبان فارسی وجود داشته باشند که املای یک واژه را متفاوت بنویسند، سامانه‌های رایانه‌ای مشکل پردازشی در متون فارسی خواهند داشت. تعدد واژه

هایی کـه مسـتعد چنـدگانگی در نوشـتار هسـتند، مـلاک تعیـین اهمیـت اشکالات دستور خط فارسی در یکسان سازی چهره‌ی خط بـرای کـاربران رایانـه اسـت. در زمینه قواعدی که دسـت کـاربران را بـرای نگـارش آزادانـه‌ی واژه‌ها مـی‌بنـدد، در بسیاری از محافل ادبی بحث‌های فراوانی شده است، همانطور که نگرش‌های جدانویسی و پیوسته نویسی افراطی در برهه‌های زمانی متفاوت، رد یا تایید شـدند. اما نبایـد ازیـاد بـرد کـه پیشـرفت کشـور در حـوزه‌ی فـنآوری اطلاعـات، در گـرو اهمیـت دادن بـه مسـئله‌ی یکسـان‌سـازی چهـره‌ی خـط فارسـی اسـت. نبـود تحـول در ایـن حـوزه و نداشـتن تمایـل سازمان‌های رسمی و غیررسمی به تقبل پروژه‌هایی که خاص زبان فارسـی باشـند، ناشـی از اطلاع آن‌ها از به هم ریختگی قواعد دستور خط فارسی است. در واقـع، اگـر خـط فارسـی و قواعد نگارش تا این حد ابهام‌زا نبود، ما نیز کمی پس از دهـه ی ۱۹۸۰ کـه اولـین خطـایـاب املایی تجاری زبان انگلیسی به بازار عرضه گردید، می‌توانستیم این سامانه را در ایران تولید نماییم. ولی مشکلات خط فارسی از یک سو، و مشکلات زبان فارسی و فارسی‌سازی سیستم‌عامل‌ها به صورت‌های گوناگون از سوی دیگر، موجب شده‌اند که مشکل زبان فارسی برای رایانه دو چندان شود.

در زمینه‌ی فاصله گذاری میان واژه‌ها کـه یکـی از بـزرگتـرین چـالش‌هـای کنـونی اشکالات پردازشی دستور خط فارسی است، می‌توان دلایل گروه‌های موافق با پیوسته نویسی و جدانویسی را بـه این صورت دسته‌بندی نمود: پیوسته نویسی کامل درج فاصله برای کاربرانی که می خواهنـد واژه‌های ماننـد «میشـود» را بـه صـورت «می‌شود» بنویسند باعث خواهد شد کـه «مـی» در مـواردی در انتهـای خـط اول باقی بماند و «شود» به ابتدای سطر بعد منتقل گردد. ایـن کـار از خوانـایی نوشـته می کاهد.اگر برای جبران انتقـال «مـی» بـه سـطر بعـد، از نـیم فاصـله اسـتفاده شـود مسـتلزم آشنایی کاربران با این کلید و موقعیت آن در صفحه کلید اسـت. زدن ایـن کلیـد تنها برای کاربرانی ساده است که با گذشت زمان به آن عادت کرده‌اند. اگر واژه‌ها پیوسته نوشته شوند و

واژه نامه‌ها نیز بر اساس اصول پیوسته نویسی تدوین شده باشند، جستجوی واژه‌ها در واژه نامه بسیار آسان خواهد بود. در فرآیند جدانویسی کامل، چشم انسان واژه‌های پر حرف را به درستی تشخیص نمی‌دهد. مثلاً اگر بخواهیم واژه «عافیت‌طلب» را به شکل «عافیتطلب» بنویسیم، خواندن آن برای هر خواننده‌ای مشکل است و نیاز به مکث اضافه دارد. در نتیجه جدانویسی منجر به سادگی خواندن واژه‌ها می‌گردد. اگر تمام واژه‌ها جدا نوشته شوند، ابهامی در معنای کلام باقی نمی ماند. هر بخش از واژه به صورت مجزا نوشته می‌شود. در حالی که در پیوسته نویسی، مشخص نیست که تا کجا باید واژه‌ها را به هم متصل نمود و اجزای واژه دقیقاً کدامیک هستند. جدانویسی در صورتی که همراه با استفاده از نیم فاصله[1] باشد، فرایند تفکیک واژه‌ها را ساده می کند. زبان‌های طبیعی که اجزای متصل کمتری دارند، در پردازش رایانه‌ای ساده‌تر هستند. به عکس هر چه بخش‌های بیشتری از واژه به هم متصل شوند، تشخیص اجزا برای رایانه پیچیده تر خواهد شد. در نتیجه، مواردی همچون تشخیص و تصحیح املاء واژه و تشخیص نقش دستوری آن مشکل تر خواهد شد (Mahboubi et al., 2017). راه حل بینابینی نه کاملاً پیوسته و نه کاملاً جدا نویسی است. به این ترتیب ظاهر واژه‌ها واژه حفظ می‌شود و حداقلی از اصول اولیه نیز رعایت می‌گردد. دستورالعمل‌های مجزا برای جدانویسی و پیوسته نویسی ارائه می‌شود و سایر موارد مطابق با سلیقه‌ی نویسنده خواهد بود. با وجود رویکردهای مختلف، نکاتی وجود دارند که در جمع بندی نهایی نباید فراموش شود. اگر حروف واژه زیاد شوند، پیوسته نویسی واژه آن را از شکل قابل خواندن خارج می‌نماید. خواندن واژه‌هایی که بیشتر از ۷ الی ۸ حرف در بخش اصلی خود دارند، برای بیشتر خوانندگان سخت و نیازمند به مکث است. چشم خوانندگان به الگوی واژه بیش از حروف آن‌ها عادت دارد. در واقع بسیاری از واژه قبل از آن که در ذهن انسان حرف به حرف پردازش شوند،

[1] Half space or pseudo space

بـه کمـک شکل شـان در ذهن تشخیص داده می‌شوند. مثلاً واژه‌ای دارای غلط املائی مانند«خدافاحظ» با آن کـه اشتباه نوشتاری است، در ذهن خواننده «خداحافظ» را تداعی می نمایـد. در حـالی کـه اگـر حـرف بـه حـرف پـردازش شود، چنین برداشتی از واژه نخواهیم داشت. بـا توجـه بـه ایـن واقعیـت، هر دسـتورالعملی کـه بخواهد شکل بسیاری از واژه‌ها را تغییر دهد، با مقاومت‌هایی روبرو خواهد شد (Taghavi et al., 2005).

اگر از جدانویسی واژه‌ها بی‌رویه استفاده شود، منجر به اشکال در خواندن واژه خواهـد شد. مثلاً اگر «بدبختانه» به صورت «بد بختانه» نوشته شود، خوانـدن آن دشـوار و بـه چشـم بیننـده ناآشـنا خواهـد آمـد. بایـد دقـت داشـت کـه برخـی از اجـزای واژه ماننـد بسـیاری از پسوندها، مدت‌هاست که در واژگان زبان نفوذ کرده‌اند و جایگاه محکمی یافته‌اند. واژگـان تولیدی از این راه، با وجـود آن کـه در ذات خـود مرکـب هسـتند امـا بـدون پسـوند خـود، کاربردی متفاوت خواهند یافت و از دید خوانندگان هم پیچیده به نظر خواهند رسید. با وجود آن که نباید شکل واژگان زبان را با یک قاعده زیر و رو نمـود، امـا بسـیاری از واژگان برای چشم‌های کاربران، متفاوت به نظر می‌رسند. برای مثال، بسیاری از کـــاربران بـه واژه «می‌شود» عـادت کرده‌اند در حـالی کـه گـروه عمـده ی دیگـری چنـین نیسـتند. بـرای یکسان سازی چهره‌ی خط فارسی، در نهایت تغییر برخی از روش‌های نگارش امری نـاگزیر خواهد بود. بزرگترین مشکل راه حل‌های بینابینی، ابهام آن‌ها در اسـتفاده از قواعـد اسـت. زمانی که انتخاب شـکل نگـارش واژه رسـماً بـه سـلیقه‌ی نویسـنده سـپرده شـود، اشـکالات پردازشی زبان آغاز خواهند شد.

برای روشن شدن عمق تاثیر خط فارسـی، نمونه‌هایی از کاربردهایی که نیاز به قاعده‌مند شدن زبان فارسی دارند، به طور خلاصه فهرست می‌شود:

- خطایابی املایی

- خطایابی ویرایشی و دستوری
- موتور جستجوی فارسی
- بازشناسی خودکار حروف فارسی
- خلاصه سازی فارسی
- استخراج واژه‌های کلیدی متن
- شباهت سنجی میان متون
- پالایش متون
- ترجمه ماشینی
- دسته بندی و خوشه بندی متون
- انواع داده کاوی و متن کاوی
- نمایه سازی

تمامی کاربردهای یاد شده، نیازمند تدوین قوانین مدون، یکسان سازی نگارش واژه‌ها، تولید ابزارها و پیکره ها و به طور اخص انجام پژوهش‌های بنیادی در حوزه‌ی پردازش زبان طبیعی برای زبان فارسی هستند. بنابراین تولید پیکره‌ها و ابزارها از جمله پیشنیازهای پژوهش در خصوص پردازش زبان طبیعی است.

ستوده و همکارانش (۱۳۹۱) در پژوهشی چالش‌های خط فارسی در رایانه را بررسی کرده‌اند و برخی از چالش‌های این حوزه را مورد بررسی قرار داده‌اند که در جدول ۱-۱ نشان داده شده است. حتی در خصوص چالش‌های ذکر شده نیز زبان‌شناسان نظرهای متفاوتی دارند و اتفاق نظر در خصوص راه حل این چالش‌ها وجود ندارد.

جدول ۱-۱-چالش‌های خط فارسی در نگارش رایانه‌ای (ستوده، ۱۳۹۱)

ردیف	چالش	ردیف	چالش
۱	تشدید (معیّن/ معین)	۲۳	گوناگونی معادلهای علمی
۲	همزۀ پایانی (املاء/ املا)	۲۴	(عدم)استفاده از «ء» بعد از« های» بیان حرکت در حالت مضاف (خانۀ مردم / خانه مردم(
۳	تنوع شیوۀ دگرنویسی (امریکا / آمریکا)	۲۵	تنوع نگارش یای وحدت نکره بعد از «های» مختفی (خانه‌ایی / خانه‌یی/ خانۀ)
۴	های غیر ملفوظ (مورچگان/مورچه‌گان)	۲۶	عدم تمایز حروف بزرگ و کوچک در ابتدای جمله
۵	همزۀ متصل به «یای» وحدت (عطایی/ عطائی)	۲۷	شباهت اعداد (صفر و نقطه / ۱ و ۲ و ۳)
۶	استفاده از «آ» و «ا» به جای هم (درآمد/ درامد)	۲۸	تعدد حروف دندانه‌دار (پیشینیان)
۷	تنوع حروف (اطاق/ اتاق)	۲۹	تعدد نقطه‌های حروف (ث ش پ)
۸	الف کوتاه (تقوی/ تقوا)	۳۰	شباهت شکل حروف (ک گ / ت ث / ر ز)
۹	تای نقطه‌دار (مشکوه/ مشکات / مشکوۀ)	۳۱	ناتوانی در نشان دادن تلفظ‌های باستانی و میانه، گویشها و لهجه‌ها
۱۰	«ی» صامت میانجی (پرتوی آفتاب/ پرتو آفتاب)	۳۲	یکسانی نشانۀ واژۀ بستهای ربطی فعل «بودن» و «م» مالکیت (پدرم =پدر من / پدر هستم)
۱۱	خط تیره (اقتصادی اجتماعی/ اقتصادی- اجتماعی)	۳۳	یکسانی علامت نکره و اسم ساز و صفت ساز (اجتماعی: اجتماع+ی نکره؛ اجتماعی بودن)
۱۲	نقطه در سرنام‌ها (اچ. آی. وی/ اچ‌آی‌وی)	۳۴	آرایش آزاد سازه‌های جمله (دیروز من کتاب خریدم/ من دیروز کتاب خریدم)
۱۳	پیوسته‌نویسی (سرهم یا با نیم‌فاصله) یا جدا نویسی (کتاب شناسی / کتابشناسی/ کتاب‌شناسی)	۳۵	فقدان پایانه‌های تصریفی نمایانگر حالت کلمه در جمله (این کار- خانه را خراب کرد. این کارخانه- را خراب کرد. این- کارخانه را خراب کرد.)
۱۴	تنوع نشانه‌های جمع (عاقلان/ عقلا / عاقلها)	۳۶	اختیاری بودن فاعل ([علی] به مدرسه رفت)
۱۵	تنوین (واقعا/واقعاً/ واقعن)	۳۷	اشتقاق صفر و تغییر مقولۀ واژگانی کلمه‌ها (انتخابها در شرایطی بد بود/ بد و خوب را تشخیص داد.)
۱۶	فاصلۀ بین حروف یک واژه به اشتباه یا به عمد (دوا زده/ دوازده؛ کدگذاری/کد گذاری)	۳۸	واژه‌های به وام گرفته یا ترجمه شده (کامپیوتر/ رایانه)
۱۷	املاهای مختلف همزه (مسئول/ مسؤول)	۳۹	مترادف‌ها (درست/ صحیح)

۱۸	تفاوت در آوا / اعراب (مَرد/ مُرد، دیر (زمان) / دیر [صومعه])	۴۰	اسامی‌عامیانه، تجاری، مشهور یا علمی
۱۹	تعدد شکلهای یک حرف (عـ ـعـ ـع ع)	۴۱	کسرهٔ اضافه (پدر او را تحسین کرد/ پدرِ او را تحسین کرد)
۲۰	یکسانی تلفظ برخی حروف (س ص ث)	۴۲	آوانویسی به جای ترجمه (سورس/ منبع)
۲۱	نوشتن «ک» و «گ» با سرکش و بی آن (ک/ ک)	۴۳	همنام‌ها و هم‌آواها شیر (ماده نوشیدنی، حیوان، ابزار)
۲۲	نگارش از راست به چپ		

۱-۳-ضرورت و اهمیّت پژوهش

پژوهش‌های زبان شناسی در حیطه‌ی زبان فارسی، نیازمند ابزارها و پیکره‌ها است. اگر ابزارها و پیکره‌ها در اختیار پژوهشگران نباشد، پیش‌نیازهای پژوهش برای علاقه‌مندان به حوزه‌ی زبان‌شناسی فراهم نخواهد شد. بنابراین پژوهش‌های بنیادی که منجر به ساخت پیکره و ابزارهای با کیفیت برای استفاده‌ی پژوهشگران زبان فارسی گردد، از اهمیت ویژه‌ای برخوردار است.

در این میان یکی از بنیادی‌ترین ابزارهای مورد نیاز در هر پژوهش زبان‌شناسی مقسم جمله به واژه و یا تجزیه‌کننده است. وظیفه‌ی این ابزار تجزیه کردن متن به واژه است. با توجه به چالش‌های زبان فارسی، تشخیص محدوده‌ی واژگان خود یک چالش بزرگ است و باید این مشکل با کمک روش‌های آماری مبتنی بر پیکره و همچنین با استفاده از قواعد دستوری زبان فارسی مرتفع گردد.

گام نخست در پردازش متون، تجزیه‌ی متن و تقسیم متن به واژه است. در زبان‌هایی مانند انگلیسی علاوه

برآنکه چنین ابزاری به راحتی در دسترس است، قواعد نگارش نیز به صورتی است که برای درصد بالایی از واژگان کاراکتر فاصله[1] جداساز واژه است (Taghavi, Young, Coombs, Pedra, Beckey, & Sadeh, 2003). برای زبان فارسی، به دلیل نوع نگارش متن فارسی در رایانه، عدم وجود قواعد مدون و عدم وجود ابزارهای زبان‌شناسی واژگان فارسی وقتی به صورت الکترونیکی نگارش می‌شوند به صور گوناگون ظاهر می‌شوند (مثال در جدول ۱). چالش دیگر آن است که در زبان فارسی صرفا کاراکتر فاصله جداساز واژگان نیست و عملاً بدلیل عدم استفاده‌ی صحیح از کاراکتر نیم فاصله در فارسی و زبان‌های مشابه مثل عربی، جداسازی واژگان چالشی بزرگ است. بنابراین برای جداسازی صحیح واژگان نیاز به بکارگیری روش‌های ترکیبی مانند روش‌های مبتنی بر پیکره و قانون است. عدم شناسایی صحیح محدوده‌ی واژه می تواند موجب کاهش دقت برنامه‌های کاربردی در این حوزه مانند غلط‌یاب‌ها و ... گردد.

جدول ۱-۲- مثالی از نحوه‌ی نگارش واژگان فارسی در رایانه

صورت اول واژه	صورت دوم واژه
کتاب<فاصله>ها	کتاب(نیم‌فاصله)ها
می <فاصله>روم	می(نیم‌فاصله)روم
آب<فاصله> دار	آبدار (سرهم)

همانطور که در جدول ۱-۲ نیز نشان داده شد عدم وجود قواعد مدون و ابزارهای رسمی برای نگارش متن موجب تولید متون الکترونیکی به تمامی صورتهای ممکن می‌شود بنابراین نیاز است که جمله به شکل صحیحی به واژگان شکسته شود. همچنین لازم است واژگان مرکب شناسایی شود. با توجه به اینکه واژگان مرکب دارای بسامد بالایی در زبان فارسی است شناسایی واژگان مرکب خود چالش بزرگی است و عدم شناسایی صحیح واژگان

[1] Space character (ASCII code 32)

مرکب موجب کاهش دقت تجزیه کننده می‌شود. باید توجه داشت پردازش زبان طبیعی، برای بکارگیری دانش نهفته در متون رقمی و پردازش آن بسیار حائز اهمیت است. برای بسیاری از زبان‌های متداول مانند انگلیسی، آلمانی و فرانسه ابزارهای پیش‌پردازش و پردازش زبان طبیعی موجود است و به سهولت در اختیار پژوهشگران قرار می‌گیرد. زبان فارسی علیرغم مشکلاتی که در تدوین استانداردهای بکارگیری در رایانه دارد، دارای مشکل در دسترس نبودن منابع و ابزارهای زبانی است. ابزارهای پردازش زبان، توابع و قطعه کدهایی هستند که در زمان پیش‌پردازش و یا پردازش زبان، بکار گرفته می‌شوند تا مصادیق پردازش‌های زبانی را ممکن سازند. برای مثال ابزارهایی مانند تجزیه کننده‌ی متن، ریشه‌یاب و غلط‌یاب متن را می‌توان از جمله ابزارهای پردازش زبان دانست. منابع زبانی شامل پیکره‌ها، داده‌ها و پایگاه‌های اطلاعاتی هستند که به صورت الکترونیکی قابل پردازش هستند و در پژوهشگران حوزه‌ی پردازش زبان طبیعی می‌توانند از این منابع استفاده نمایند. برای نمونه می‌توان از پیکره‌های متنی زبانی، پیکره واژگان، گنجینه لغات، اصطلاحنامه‌ها و دیکشنری‌ها به عنوان منابع پرکاربرد نام برد.

پیش‌نیاز ساخت بسیاری از برنامه‌های کاربردی در هر زبان، نیازمند وجود ابزارها و منابع زبانی است (Shamsfard et al., 2010; Shamsfard, 2011). همانطور که در زبان‌های رایج بکارگیری این منابع و ابزارها موجب تولید نرم‌افزارهای کاربردی مفیدی مانند غلط‌یاب، شناسایی نوری نویسه‌ها (OCR)، مترجم متن و خلاصه ساز متن شده است، نیاز به این نرم‌افزارهای کاربردی در زبان فارسی نیز حس می‌شود. به هر ترتیب پیش‌نیاز ساخت این کاربردها که هم اکنون از جمله طرح‌های ملی هستند، انجام پژوهش در زمینه گسترش ابزارها و منابع زبان فارسی است.

تجزیه کننده‌ی متن ابزاری است که با دریافت متن، آن را به اجزای تشکیل دهنده، (واژه‌ها) تجزیه می‌کند. باید توجه داشت که برای بسیاری از زبان‌ها مانند انگلیسی ساخت چنین ابزاری پیچیدگی‌های زیادی

ندارد، زیرا مقسم واژه در زبان صرفا کاراکتر فاصله و تعدادی کاراکترهای کنترلی و نشانه‌گذاری است (Taghavi et al., 2003). در زبان فارسی مقسم واژگان کاراکتر فاصله نیست و در نظر گرفتن آن به تنهایی نتیجه‌ی موثری را در پی ندارد. زبان فارسی دارای نویسه‌هایی است که به صور گوناگون نوشته می‌شوند. این نویسه‌ها با توجه به اینکه در ابتدا، میان و یا در انتهای واژه قرار گیرند، شکل متفاوتی از نظر نوشتاری دارند (Dastgheib, Fakhrahmad, & ZolghadriJahromi, 2016). باید توجه داشت که قوانین نگارش نویسه‌های فارسی (بخصوص در رایانه) مدون نیست و تا کنون چندین باز از پیوسته‌نویسی به گسسته‌نویسی تغییر کرده است(Dastgheib et al., 2016). دانش آموختگان زبان فارسی که به اصول نگارش فارسی مسلط هستند در نگارش خط فارسی مشکل دارند و اتفاق نظری برای جدا و یا پیوسته نوشتن واژه‌های مرکّب وجود ندارد (Dastgheib et al., 2016; Shamsfard, 2011). حال آنکه با وجود چنین مشکلاتی وقتی خط فارسی در رایانه الکترونیکی تایپ می‌شود، مشکلات عدیده خط فارسی دو چندان می‌شود زیرا عدم وجود ابزارهای مناسب پردازش زبان موجب می‌شود کاربر رایانه نتواند متن فارسی را صحیح نگارش کند. به عنوان مثال نشانه‌ی جمع فارسی (ها) می تواند به صورت متصل و یا جدا نوشته شود. جدول ۱-۳ نحوه‌ی نگارش صور مختلف یک واژه را نشان می‌دهد.

جدول ۱-۳-صور مختلف نوشتن شکل جمع واژه‌ی «آب» در رایانه

آبها (متصل)
آب‌ها (جدا با نیم‌فاصله)
آب ها (جدا با فاصله)

همانطور که در جدول ۱-۳ نشان داده شده است، جمع واژه‌ی آب به سه صورت نوشته شده است و تمامی

حالات صحیح است و باید توسط تجزیه کننده‌ی متن به درستی شناسایی شود. علاوه بر آن واژه‌های مرکّب در فارسی بسیار پر کاربردند و بیش از نیمی از واژگان را شامل می‌شوند. بنابراین بکارگیری یک تجزیه کننده‌ی متن مبتنی بر دانش و نه فقط قوانین نگارش فارسی می تواند دقّت بالایی داشته باشد. این ابزار باید برای استفاده پژوهشگران منتشر شده و در اختیار علاقه‌مندان باشد.

وجود پیکره‌های زبانی نیز از جمله منابع مهم زبان فارسی است که کمبود آن موجب خسارت‌های زیادی برای زبان فارسی است. یکی از منابع زبانی مورد نیاز برای هر زبان، مجموعه‌ی واژگان متداول زبان است. این واژگان را نمی توان از دیکشنری‌ها استخراج نمود زیرا بسیاری از واژه‌های نامتداول نیز در آن وجود دارد.

پیکره‌ی مهم دیگر، مجموعه‌ی واژگان دارای ابهام در فاصله ویرایشی مشخص نسبت به هر واژه است. به عبارت دیگر واژگان صحیح فارسی که در فاصله ویرایشی مشخصی (مثلاً یک) نسبت به واژه‌ی هدف قرار دارند مجموعه‌ی ابهام این واژه نامیده می‌شوند. این مجموعه، مجموعه‌ی ابهام[۱] نام دارد. ساخت چنین پیکره‌ای مسلماً بسیار زمانبر است و هزینه‌ی پردازشی بالایی دارد و باید بصورت برون خطی[۲] انجام شود. کاربرد این پیکره در شناسایی نوری نویسه‌ها و غلط‌یابی متن است. از جمله دستاوردهای دیگر این پژوهش تولید پیکره‌ی مجموعه‌ی ابهام برای واژگان فارسی در فاصله‌ی ویرایشی یک از واژه‌ی هدف است. این مجموعه در کاربردهایی مانند شناسایی نوری نویسه‌ها، غلط‌یاب و ... کاربرد دارد. با توجه به اینکه کاربردهای این پیکره بصورت برخط است، تولید آن در زمان اجرا موجب کندی سرعت عملکرد برنامه کاربردی می‌شود بنابراین بهتر است بصورت برون خطی محاسبه شده و چنین پیکره ای در دسترس باشد. به عنوان مثال برای واژه‌ی برق برخی از واژگان مجموعه‌ی ابهام عبارتست

[۱] Confusion Set
[۲] Off-line

از: ابرق، بارق، برحق، مبرق، براق و ... همانطور که در مثال مشاهده می‌شود، این واژه‌ها، واژه‌های صحیح فارسی هستند و در فاصله‌ی ویرایشی یک از واژه‌ی برق قرار دارند، بنابراین با اعمال یک اپراتور ویرایشی به‌روی این واژه‌ها، واژه‌ی هدف (برق) بدست می‌آید.

هدف این پژوهش در راستای رفع محدودیت منابع و ابزارهای زبان شناسی برای زبان فارسی است. لذا در این راستا به تولید پیکره مجموعه ابهام واژه های فارسی که در فاصله ویرایشی یک قرار دارد و ابزار تجزیه کننده جمله به کلمه (واژه) تولید می شود. بدون تردید هر گامی که در راستای تهیه ابزارها و منابع برای زبان فارسی برداشته شود می تواند پژوهشگران این حوزه را یاری نماید.

فصل دوم

مروری بر پژوهش‌های انجام شده

۲-مروری بر پژوهش‌های انجام شده

۲-۱-مقدمه

ساخت منابع زبانی در دنیا برای زبان‌های رایج از دیرباز مرسوم بوده و بصورت متن باز و یا توابع در اختیار پژوهشگران قرار می‌گیرد. در این میان آزمایشگاه پردازش زبان دانشگاه استنفورد در این زمینه پیشگام است و ابزارهای پردازش زبانی برای زبان انگلیسی را به تفضیل و رایگان رسماً منتشر نموده است[1] .

به هر ترتیب با توجه به نیاز به پیکره‌ها برای انجام پژوهش‌های پیکره محور و قواعد خاص زبان فارسی، لازم است منابع زبانی برای زبان فارسی بومی‌سازی شود. استفاده از منابع زبان‌های دیگر برای زبان فارسی مشکلات عدیده‌ای منجمله عدم همخوانی نتایج را در برخواهد داشت. در ادامه به مهمترین پژوهش‌های انجام شده در حوزه‌ی زبان فارسی پرداخته خواهد شد.

۲-۲-تاریخچه پژوهش‌های انجام شده

پردازش زبان طبیعی را می‌توان در سطوح مختلف زبان بسته به نیاز به انجام رسانید. پردازش زبان در هر سطح، نیازمند دانش، منابع و پیکره‌های مورد نیاز آن سطح و سطوح پایین‌تر است. در دسترس بودن منابع و دانش برای انجام تحقیق در حیطه‌ی پردازش زبان طبیعی از جمله چالش‌های پردازش زبان طبیعی است (Dastgheib et al., 2016).

ساخت منابع زبانی و ابزارهای مورد نیاز پردازش زبان فارسی، با گسترش استفاده از دستگاه‌های

[1] https://nlp.stanford.edu/software/

الکترونیکی افزایش یافته است. در این میان برخی از پیکره‌ها و ابزارهای حاصل شده از پژوهش منتشر شده است ولی بیشتر این ابزارها یا رسماً منتشر نشده است و یا در اختیار پژوهشگران نیست (Dastgheib et al., 2016;). در ادامه برخی از مهمترین پژوهش‌های اخیر که در راستای توسعه‌ی زبان فارسی انجام شده است، مرور می‌گردد.

موسوی و همکاراش (۲۰۱۷) از روش یادگیری بدون نظارت برای تولید اصطلاحنامه برای زبان فارسی استفاده کرده‌اند. در این پژوهش با استفاده از یک پیکره متنی فارسی و دیکشنری دو زبانه، مجموعه‌های واژگان وابسته استخراج و در نهایت اصطلاحنامه تولید شده است. برای این اصلاحنامه دقت ۹۱٪ و ۱۶۰۰۰ واژه گزارش شده است (Mousavi & Faili, 2017). در پژوهش مشابهی شمس‌فرد (۲۰۰۸) اقدام به تهیه‌ی اصطلاحنامه برای زبان فارسی نمود، نتیجه‌ی این پژوهش به نام اصطلاحنامه‌ی فارس‌نت در دسترس پژوهشگران قرار گرفته است (Shamsfard, 2008). هم‌اکنون نسخه‌ی دوم این اصطلاح‌نامه نیز منتشر شده است.

محبوبی و همکارانش (۲۰۱۷) مدلی را برای پردازش متون زبان فارسی ابداء کرده‌اند که به عنوان اختراع نیز آن را به ثبت رسانده‌اند. در این پژوهش، یک مدل برای پردازش و برچسب زدن موقت متون فارسی ایجاد شده است و در پژوهش به اهمیت وجود ابزارها و تشخیص محدوده‌ی واژه‌های فارسی[۱] نیز اشاره شده است (Mahboubi et al., 2017).

تهیه ابزارهای زبانی برای پردازش زبان فارسی نیز بسیار مهم و از جمله پیش‌نیازهای این حوزه است. اصغر و همکارانش (۲۰۱۳) در پروژه‌ای برای زبان فارسی مشکلات پردازش متن را برشمردند و یک تجزیه کننده‌ی متن و ابزارهایی برای پیش‌پردازش مانند حذف کلمات زائد را ارائه نمودند. نتیجه این پژوهش بصورت رسمی

[۱] Tokenizer

منتشر نشده است (Asghar, Khan, Ahmad, & Kundi, 2013).

سراجی در پژوهشی سعی در تهیه ابزارهای پردازش زبان برای زبان فارسی نموده است. این مجموعه ابزار به نام SETPer با استفاده از قوانین و مبتنی بر پیکره‌ها سعی در شناسایی محدوده‌ی واژگان دارد (Seraji, Megyesi, & Nivre, 2012). این ابزار برای نرم افزار تحلیل پیکره‌ی Uplug طراحی شده است[۱]. در ضمن این ابزار دارای دو قسمت یکی برای بخش بندی مجزای جمله و دیگری برای تجزیه‌ی جمله به واژه است. مشکل اصلی این پژوهش عدم در نظر گرفتن نیم‌فاصله برای برخی واژگان فارسی است.

شمس فرد (Shamsfard et al., 2010) در پژوهشی در این حوزه اقدام به تهیه برخی از ابزارهای مورد نیاز برای پردازش زبان طبیعی فارسی نموده است. این مجموعه شامل تجزیه کننده‌ی متن، غلط‌یاب، تحلیل‌گر زبان‌شناسی و برچسب‌گذار نقش واژگان است. در این ابزار با تکیه بر شناسایی مفصل‌های زبان فارسی، علامت‌ها و نشانه گذاری‌ها و شکل الفبای فارسی سعی شده است که محدوده‌ی واژگان را شناسایی نماید. برخی از ابزارهای این مجموعه بصورت وب پایه ارائه شده است[۲] و بقیه ابزارها رسماً منتشر نشده است[۳]. مشکل ابزارهای وب‌پایه آن است که بصورت جاسازی شده[۴] امکان استفاده از آن وجود ندارد.

منصوری راد و همکارانش (۲۰۰۰) در پژوهشی به نام پروژه‌ی شیراز سعی درتولید ماشین ترجمه برای زبان فارسی نمودند. در این پژوهش با توجه به نیاز به پردازش و پیش پردازش متون فارسی یک تجزیه‌کننده‌ی متن بصورت اختصاصی بنام Posttokenizer تولید شده است. این تجزیه‌کننده‌ی فارسی در فاز پیش پردازش

[۱] http://stp.lingfil.uu.se/~mojgan/setper.html
[۲] http://step1.nlplab.sbu.ac.ir
[۳] برای مشاهده مثال به ضمیمه ۱ مراجعه فرمایید
[۴] Embeded

متون فارسی بکارگرفته شده است (Amtrup, Rad, Megerdoomian, & Zajac, 2000). تجزیه کننده به عنوان یک محصول مستقل منتشر نشده است. دسترسی به ابزارهای پژوهش فوق نیز برای پژوهشگران امکان پذیر نمی‌باشد.

آصف پور معمومی (۱۹۹۰) در آزمایشگاه فناوری‌های وب دانشگاه فردوسی مشهد اقدام به تهیه ابزارهای پردازش زبان نموده است که بخشی از آن به تجزیه کننده‌ی متن اختصاص دارد. این تجزیه کننده براساس عبارات قاعده‌مند کار می‌کند و برای محیط متن باز گیت[۱] نوشته شده است. این بسته بصورت کد بسته در اختیار پژوهشگران است. استفاده از عبارات قاعده‌مند بدلیل استثناهای زیادی که در فارسی وجود دارد، کارآیی پایین‌تری نسبت به سیستم‌های آماری و برپایه‌ی دانش دارد.

پیکره حجم زیادی از داده‌های زبانی است، که بر اساس معیارهای مشخص، برای هدف معینی جمع آوری و ذخیره شده است، بطوری که نماینده زبان یا گویش مورد مطالعه باشد. به طور کلی در طراحی و تهیه یک پیکره یکی از مهمترین مسائلی که باید مورد توجه قرار بگیرد، هدف غائی پیکره و منظوری که از پیکره مد نظر است می‌باشد. پیکره‌ها از نظر تعداد زبان، به دو دسته‌ی یک زبانه و یا چند زبانه تقسیم می‌شوند. پیکره‌های چند زبانه دارای متن در بیش ازیک زبان هستند (Dastgheib et al., 2016; Eghbalzadeh, Hosseini, Khadivi, & Khodabakhsh, 2012; Shamsfard et al., 2010).

متاسفانه تاکنون پژوهشی در خصوص تولید پیکره مجموعه‌ی ابهام واژگان فارسی تهیه و ارائه نشده است و بصورت رسمی نیز منتشر نشده است (Dastgheib et al., 2016). هدف این پژوهش تولید چنین پیکره‌ای

[۱] Gate open source

است که مجموعه‌ی ابهام واژه‌های فارسی را دربر داشته باشد. برخی از تلاش‌های ذکر شده در خصوص تولید تجزیه‌کننده‌ی متن فارسی نیز یا رسماً منتشر نشده است و یا بدلیل مشکلات ساختاری فارسی برای حوزه‌ی محدودی تهیه شده و بصورت ابزار ارائه نشده است. بنابراین بصورت جاسازی شده در نرم‌افزارها یا پژوهش‌ها قابل استفاده نیست. نسخه‌های دیگر ارائه شده هم معمولاً قابل بکارگیری در ساختارهای تولید نرم افزار کارآیی ندارد و همچنین بعضاً بدلیل بکارگیری ساختارهایی مثل وب‌پایه قابل تزریق و چسباندن به نرم افزار کاربردی نیست و دارای مشکلاتی مانند سرعت کم، دقت نامناسب، واژگان خارج از دیکشنری و ... است و امکان تنظیم آن برای کاربرد خاص نیز وجود ندارد. در حال حاضر بسیاری از دانشگاه ها و موسسات پژوهشی در حال تحقیق در حوزه‌ی زبان‌شناسی هستند و برای این حوزه ابزارهایی ارائه می‌کنند. از این جمله می توان به دانشگاه فردوسی مشهد و دانشگاه صنعتی شریف اشاره کرد. علیرغم ارزشمند بودن پژوهش‌های این حوزه باید خاطر نشان کرد بدلیل مشکلات فوق الذکر لازم است ابزاری کاربردی با رویکرد قابلیت کاربرد در تولید نرم افزار تولید شود تا بتوان از آن برای بالا بردن کیفیت نرم افزارها استفاده کرد. به عنوان مثال هم‌اکنون مرکز منطقه‌ای اطلاع رسانی علوم و فناوری اقدام به تولید سامانه‌ی ژورنال‌یاب[1] نموده است. برای افزایش سرعت و دقت این سامانه و بسیاری کاربردهای دیگر در حوزه‌ی زبان‌شناسی و کاربرد ابزارهایی مانند تجزیه‌کننده در مطالعات دانشجویی تولید چنین ابزاری ضروری بنظر می‌رسد. نمونه های مشابه یاد شده دارای مشکلات سرعت، دقت، محدود بودن حوزه، در دسترس نبودن بصورت متن باز یا ابزار (DLL) هستند و بنابراین نیاز است چنین ابزاری برای بکارگیری در پژوهش‌های مرکز و پروژه‌های دانشجویان تولید شود تا با در دسترس بودن منابع آن مانند کد، گنجینه لغت و ... بتوان آنرا برای کاربردهای نرم افزاری بهینه سازی و شخصی سازی کرد.

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

با توجه به اهمیت مجموعه‌های ابهام به عنوان یکی از منابع زبانی پژوهش‌هایی برای تولید مجموعه‌های ابهام در دیگر زبان‌ها بغیر از فارسی انجام شده است. در این پژوهش (Noaman, Sarhan, & Rashwan,2016) نوامن و همکارانش سعی در تولید مصحح خودکار لغوی زبان عربی کردند و برای انجام این مهم، مجموعه‌ی ابهام واژگان عربی را که به عنوان پیش‌نیاز در این پژوهش مورد نیاز بوده است، تولید نمودند. در پژوهشی دیگر، با استفاده از تکنیک LSTM، سعی در تولید مجموعه‌ی ابهام و رفع ابهام برای شناسایی صحبت برای زبان انگلیسی شده است. در این پژوهش نیز به اهمیت پیکره‌ی مجموعه‌ی ابهام واژه‌ها اشاره شده است (Xiong, Wu, Alleva, Droppo, Huang, & Stolcke, 2018).

در پژوهشی توسط رلو و همکارانش (۲۰۱۸)، از مجموعه‌ی ابهام برای تحت نظر گرفتن غلط‌های املایی در زبان انگلیسی استفاده شده است. این پژوهش با استفاده از تکنیک‌های یادگیری ماشین و پیکره مجموعه‌های ابهام این تحلیل ارزشمند را انجام داده است (Rello, Romero, Rauschenberger, Ali, Williams, Bigham & White, 2018,).

لازم به ذکر است علیرغم اهمیت پیکره‌ی مجموعه‌های ابهام تاکنون به صورت رسمی چنین پیکره‌ای برای زبان فارسی تهیه نشده است و در پژوهش حاضر به این مهم پرداخته شده و به عنوان محصول اصلی این پژوهش چنین پیکره‌ای در قالب XML تهیه شده و جهت استفاده پژوهشگران عرضه می‌گردد و بصورت رسمی نیز منتشر خواهد شد.

فصل سوم

روش پژوهش

۳-روش پژوهش

۳-۱-مقدمه

هدف این پژوهش تولید دو محصولِ ابزار تجزیه کننده‌ی جمله به واژه و پیکره‌ی مجموعه‌ی ابهام برای واژگان فارسی در فاصله‌ی ویرایشی یک می‌باشد. بنابراین یک ابزار برای تجزیه به واژه تولید می‌شود که پیش‌نیاز پژوهش‌ها در حوزه‌ی پردازش زبان طبیعی برای زبان فارسی است. این ابزار علاوه بر تجزیه‌ی جمله به واژه، نیم‌فاصله را در صورت نیاز اصلاح می‌کند و واژه‌ها را به شکل صحیح شناسایی می‌کند. جدول ۳-۱ مثالی از یک جمله و اصلاح نیم‌فاصله را نشان می‌دهد.

جدول ۳-۱-اصلاح نیم‌فاصله و تجزیه به واژه

جمله	واژه‌های شناسایی شده
علی<فاصله>به<فاصله>مدرسه<فاصله>می<فاصله>رود.	۱. علی ۲. به ۳. مدرسه ۴. می<نیم‌فاصله>رود

همانطور که در جدول ۳-۱ مشاهده می‌شود، علاوه بر تشخیص واژه می‌بایست واژه‌ای که اشتباهاً با فاصله از هم جدا شده است مجدداً تصحیح شود و به صورت واژه شناسایی شود. در بسیاری از متون فارسی بکار گرفتن

فاصله به جای نیم‌فاصله مرسوم است و این امر مشکلات متعددی را برای پردازش واژه‌های فارسی ایجاد می‌کند. بنابراین برای جلوگیری از این مشکل، سعی می‌شود با استفاده از گنجینه لغت، واژه‌های نادرست تصحیح شود.

محصول دیگر این پژوهش پیکره‌ی مجموعه‌ی ابهام برای واژه‌های فارسی است. پیکره‌ها منابعی ارزشمند از نظر پردازش زبان طبیعی است. مجموعه‌ی ابهام یک پیکره است که برای هر واژه، یک مجموعه از واژگان صحیح فارسی ارائه می‌کند که در فاصله ویرایشی مشخصی (معمولاً یک) از واژه‌ی اصلی قرار دارد. از جمله کاربردهای این پیکره در شناسایی نوری نویسه‌ها، غلط‌یاب‌ها و تبدیل گفتار به متن می‌باشد.

۳-۲-تجزیه کننده متن به واژه

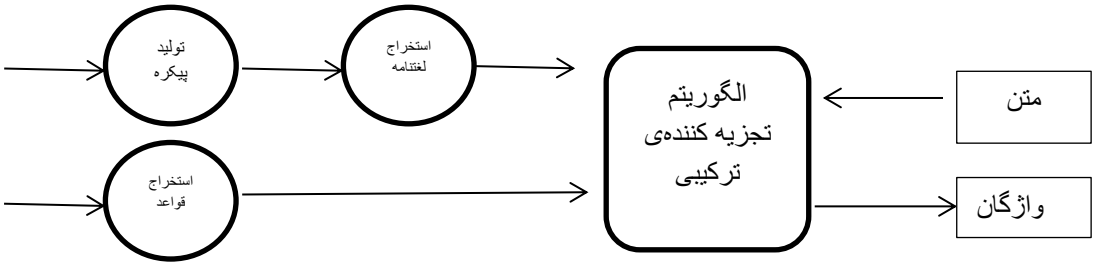
تجزیه‌کننده‌ها معمولاً براساس روش آماری مبتنی بر پیکره و یا بر اساس قانون ساخته می‌شوند. روش دیگری که برای ساخت تجزیه کننده بکار می‌رود، روش ترکیبی است. در روش ترکیبی از ترکیب قوانین مبتنی بر ساختار واژه‌های زبان و همچنین ساختار آماری مبتنی بر پیکره استفاده می‌شود.

روش‌های آماری دقت مناسبی دارند ولی جمع آوری پیکره‌ای که گستره‌ای جامع از واژه‌ها را داشته باشد، مشکل است. بنابراین چالش این روش جمع آوری دیتا برای پیکره است. مشکل دیگر نیاز به بروزرسانی‌های مکرر پیکره است، بدلیل اینکه صرفاً بر اساس پیکره و واژه‌های درون گنجینه لغت الگوریتم تجزیه انجام می‌شود، در صورتی که واژه‌ای درون پیکره وجود نداشته باشد امکان تجزیه درست آن واژه وجود ندارد، بنابراین مشکل این روش واژه‌های خارج از دیکشنری[۱] است. باید دقت داشت که تجزیه فقط بر اساس نویسه‌ی فاصله نتیجه‌ی مناسبی

[۱] Out of vocabulary words

در برنخواهد داشت و نویسه‌ی فاصله به کرات به جای نویسه‌ی نیم‌فاصله استفاده می‌شود و موجب گسستگی واژه‌های مرکب، چند بخشی و افعال می‌شود. روش مبتنی بر قانون نیز نیازمند قوانین مدون در زمینه ساختار واژه‌های زبان مورد نظر می‌باشد. تهیه قوانین مدون از جمله چالش‌های این روش است. چالش دیگر در این روش واژه‌هایی است که در مجموعه‌ی قوانین نمی‌گنجد. این واژه‌ها باید به صورت استثنا مورد بررسی قرار گیرد. بنابراین بکار گرفتن روش مبتنی بر قانون به تنهایی برای زبان‌هایی مانند فارسی که قواعد مدوّن برای ساختار واژه‌ها وجود ندارد، پاسخ مناسبی را ارائه نخواهد داد.

روش ترکیبی از ترکیب روش‌های آماری برگرفته از اطلاعات پیکره متنی و روش‌های مبتنی بر قاعده[۱]، استفاده می‌کند. این روش برای زبان‌هایی که پیچیدگی بیشتری از زبان‌های مرسوم دارند، بکار می‌رود. این روش‌ها سرعت و دقت مناسبی را می‌تواند کسب نماید. در این پژوهش از روش ترکیبی استفاده شده است. مزیت روش ترکیبی امکان تنظیم و بهینه‌سازی دقیق‌تر برای زمان پاسخ و کارآیی الگوریتم است.



شکل ۳-۱- الگوریتم تجزیه کننده

در ادامه واحدهای بکار رفته در این الگوریتم مورد بررسی قرار خواهد گرفت.

[۱] Rule

۳-۲-۱-تولید پیکره متنی

منبع تولید پیکره متن فارسی در این پژوهش، متون مقالات علمی و پیکره خبری پرسیکا (Eghbalzadeh et al., 2012) است. متن مقالات فارسی به روش نمونه‌گیری طبقه‌بندی شده تصادفی[۱] انتخاب می‌شود. در این روش مقالات در دسته‌های فنی و مهندسی، کشاورزی، هنر و معماری و پزشکی دسته‌بندی شده‌اند و از این دسته‌ها به صورت تصادفی ۱۰۰۰ مقاله استخراج شده است. جدول ۳-۲ نشان دهنده‌ی جزئیات مقالات انتخاب شده از هر گروه است.

جدول ۳-۲-مقالات انتخاب شده برای تولید پیکره

گروه	تعداد مقالات انتخاب شده
فنی و مهندسی	۳۲۴
کشاورزی	۹۸
هنر و معماری	۱۹۴
پزشکی	۳۸۴
جمع	۱۰۰۰

[۱] stratified random sampling

مقالات انتخاب شده به تمام متن تبدیل شده و پیش پردازش می‌شود. اطلاعات تصاویر و کاراکترهای اضافی از متن حذف می‌شود. متن بدست آمده برای استفاده در تولید لغتنامه ذخیره می‌شود. پس از حذف این موارد این پیکره دارای بیش از ۵ میلیون واژه و حدود ۲۴ میلیون کاراکتر است.

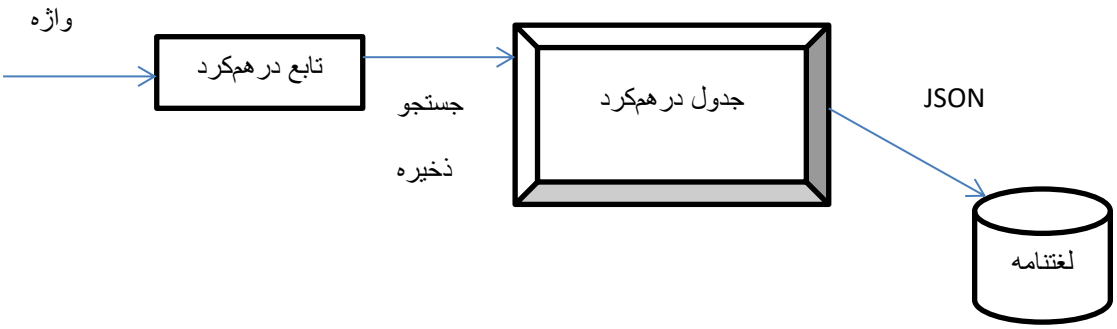
پیکره‌ی پرسیکا نیز دارای متون خبری در کلیه‌ی حوزه‌های خبری است. به دلیل آنکه متون علمی دارای دایره‌ی واژگان محدود به حوزه‌ی علمی است، مجموعه‌ی مقالات خبری که از متون روزنامه‌ها و سایت‌های خبری در این پیکره انتخاب شده است به پیکره اضافه می‌شود تا دایره‌ی لغات موجود در پیکره و لغتنامه افزایش یابد. این پیکره نیز دارای حدود ۱۱ هزار قطعه متن خبری است و دارای حدود ۴,۵ میلیون واژه و ۲۳ میلیون کاراکتر است (Eghbalzadeh et al., 2012). با توجه به اهمیت جامعیت واژه‌های استفاده شده در گنجینه، لازم است گستره و حوزه‌ی متون انتخابی مناسب باشد. بنابراین از ترکیبی از متون علمی و خبری به عنوان پیکره‌ی متنی فارسی استفاده شده است.

۳-۲-۲-تولید لغتنامه

پس از تهیه پیکره از متون مورد نظر، باید واژه‌های متون استخراج شود و با کمک آن مجموعه و محاسبه بسامد تکرار واژه‌ها، بتوان احتمال رخداد واژه‌ها را نیز محاسبه نمود. برای این منظور از ساختار دیتا دیکشنری و یا جدول در هم‌کرد[۱] استفاده می‌شود. این ساختار می تواند داده‌ها را با هزینه بسیار کم و با پیچیدگی زمانی $O(1)$ ذخیره کند. اهمیت بازیابی با سرعت بالا در زمان پاسخ الگوریتم نهایی و ساخت پیکره تاثیرگذار است. در شکل ۳-۲ الگوریتم ذخیره سازی داده‌ها (واژه‌ها) در جدول درهم‌کرد را نشان می‌دهد. همانطور که در شکل ۳-۲ مشاهده

[۱] Hash table

می‌شود، هر واژه ابتدا با محاسبه تابع درهم‌کرد در جدول جستجو می‌شود. اگر داده مورد نظر در جدول موجود باشد، فرکانس تکرار آن بروز می‌شود و در غیر اینصورت به جدول اضافه می‌شود. این ساختار در حافظه موقت نگهداری می‌شود و سپس با کمک ساختار JSON بصورت متنی در حافظه‌ی دائمی ذخیره می‌شود.



شکل ۳-۲-الگوریتم تولید لغتنامه فارسی

از جمله مزایای روش پیشنهادی امکان درج و حذف واژه در لغتنامه با استفاده از قالب JSON است. پس از اجرای الگوریتم برای تمامی واژه‌های پیکره، لغتنامه بدست آمده ذخیره می‌شود تا در مراحل بعدی با کمک آن بتوان احتمال رخداد واژه‌ها را مبتنی بر پیکره محاسبه نمود. پس از اجرای الگوریتم حدود ۱٫۸ میلیون واژه (مدخل جستجو) منحصربفرد استخراج شده و در لغتنامه ذخیره می‌شود. منظور از واژه در این لغت‌نامه، مدخل جستجو است. لذا تمام حالت‌های تصریفی واژه که در متن ظاهر شده است بنحصراً مورد پردازش قرار گرفته و در لغت نامه ذخیره شده است.

۳-۲-۳-استخراج قواعد

یکی از مشکلات زبان فارسی و زبان‌های مشابه مانند عربی، شناسایی محدوده‌ی واژه است (Kashefi et al., 2010; Noaman et al., 2017). در زبان‌هایی مثل فارسی یا عربی فقط نمی‌توان نویسه‌ی فاصله را به عنوان جداساز[۱] در نظر گرفت، زیرا این نویسه در کلمات مرکّب، پیش‌وندها و پس‌وندها استفاده می‌شود (Shamsfard et al., 2010). جدول ۳-۳ نمونه‌ای از بکارگیری نویسه‌ی فاصله در واژه‌های مرکب فارسی را نشان می‌دهد.

جدول ۳-۳-واژه‌های مرکّب در زبان فارسی

واژه با استفاده از فاصله	واژه با استفاده از نیم‌فاصله
پیش<فاصله>پردازش =پیش پردازش	پیش<نیم‌فاصله>پردازش=پیش‌پردازش
درخت<فاصله>ها =درخت ها	درخت<نیم‌فاصله>ها=درخت‌ها
می<فاصله>دوم =می دوم	می<نیم‌فاصله>دوم=می‌دوم
راه<فاصله>رفتن =راه رفتن	راه<نیم‌فاصله>رفتن=راه‌رفتن

برای تصحیح انفصال[۲] که با تایپ کردن فاصله به جای نیم‌فاصله اتفاق می‌افتد، از قاعده‌ای به صورت زیر استفاده می‌شود:

$$W^{*}=W_1+<spc>+W_2+<spc>+\ldots+W_i \qquad ۳-۱$$

[۱] Delimiter
[۲] Split-error

همانطور که در معادله‌ی ۳-۱ نشان داده شده است، یک واژه‌ی مرکب که با W^* نشان داده شده است، ممکن است با نویسه‌ی فاصله از هم جدا شده باشد. بنابراین برای تصحیح آن اجزای آن با لغتنامه بصورت آماری کنترل می‌شود، اگر بتوان واژه‌ی صحیحی را با ترکیب اجزای واژه مبهم ایجاد کرد، بصورت واژه‌ی مرکب در نظر گرفته می‌شود. به عنوان مثال راه<فاصله>رفتن دو واژه است که در لغتنامه وجود دارد، ولی بدلیل اینکه ترکیب آن یک فعل مرکب است و در لغتنامه وجود دارد (یافت می‌شود)، به عنوان یک واژه‌ی منحصربفرد شناسایی و فاصله به نیم‌فاصله تصحیح می‌شود.

1. Select a word in list as Wi
2.if Wi-1+<HS>+Wi +<HS>+Wi+1 is a correct word in dictionary then correct half space and return
3. else if Wi-1+<HS>+Wi is a correct word in dictionary then correct half space and return
4.else if Wi+<HS>+Wi+1 is a correct word in dictionary then correct half space and return

شکل ۳-۳-الگوریتم تصحیح نیم‌فاصله

در روش پیشنهادی از روش پنجره‌ای[1] بر اساس طول پنجره‌ی متغیر استفاده شده است. در این روش برای شناسایی و تصحیح درست نیم‌فاصله بر اساس گنجینه لغت واژه‌های فارسی که با استاندارد درهم‌کرد در رایانه ذخیره می‌شود، ترکیبی از واژه‌ها در پنجره با گنجینه لغت کنترل می‌شود و در صورت نیاز کاراکتر فاصله با نیم‌فاصله جایگیزین می‌شود. بنابراین اگر واژه‌ای به نادرست شناسایی شده باشد امکان تصحیح آن وجود دارد. در شکل ۳-۳ نیز نمونه‌ای از الگوریتم پنجره‌ای با طول ۲ و طول ۳ نشان داده شده است.

[1] Windowing

بنابراین برای تقویت شناخت محدوده‌ی واژه‌ها، قواعد مستخرج از ساختار واژه‌های فارسی، تدوین و در نرم‌افزار مورد استفاده قرار گرفته است. یکی دیگر از قوانین مهم در زبان فارسی، جمع بسته شدن اسامی با «ها» می‌باشد. در این قاعده، واژه به همراه علامت جمع فارسی در نظر گرفته می‌شود. اگر به اشتباه از نویسه‌ی فاصله استفاده شده باشد، با کمک این قاعده واژه اصلاح می‌گردد ومحدوده‌ی صحیح واژه شناسایی می‌شود. استفاده از قاعده‌ها با استفاده از عبارات با قاعده در نرم‌افزار به صورت ماشین متناهی مدل می‌شود. به عنوان نمونه، در ذیل برخی عبارات باقاعده برای شناسایی حروف فارسی و اعداد فارسی آورده شده است. که به ترتیب شامل شناسایی کاراکترهای فارسی، شماره موبایل در فارسی، شناسایی تاریخ فارسی، و شناسایی حروف و اعداد فارسی است. بنابراین ترکیبی از عبارات با قاعده و مبتنی بر پیکره، برای شناسایی محدوده‌ی واژه استفاده شده است.

“^([\u0600-\u06FF]+\s?)+$”

“(^(09|9)[1][1-9]\\d{7}$)|(^(09|9)[3][12456]\\d{7}$)”

“^[1-4]\d{3}\/((0?[1-6]\/((3[0-1])|([1-2][0-9])|(0?[1-9])))|((1[0-2]|(0?[7-9]))\/(30|([1-2][0-9])|(0?[1-9]))))$”

“^[\u0600-\u06ff0-9\s]+$|[\u0750-\u077f0-9\s]+$|[\ufb50-\ufc3f0-9\s]+$|[\ufe70-\ufefc0-9\s]+$|[\u06cc0-9\s]+$|[\u067e0-9\s]+$|[\u06af0-9\s]$|[\u06910-9\s]+$|^$”

۳-۳-تولید پیکره مجموعه واژه‌های مبهم فارسی (مجموعه‌های ابهام)

پیکره حجم زیادی از داده‌های زبانی است، که بر اساس معیارهای مشخص، برای هدف معینی جمع آوری و ذخیره شده است، بطوری که نماینده زبان یا گویش مورد مطالعه باشد(Dastgheib et al., 2016) . به طور کلی در طراحی و تهیه یک پیکره یکی از مهمترین مسائلی که باید مورد توجه قرار بگیرد، هدف غائی پیکره و منظوری که از پیکره مد نظر است می‌باشد. پیکره‌ها از نظر تعداد زبان، به دو دسته‌ی یک زبانه و یا چند زبانه تقسیم می شوند (Sennrich & Volk, 2010; Smith, Quirk, & Toutanova, 2010). پیکره‌های چند زبانه دارای

متن در بیش ازیک زبان هستند. پیکره‌های چند زبانه به دو دسته‌ی پیکره‌های قیاس پذیر[۱] و پیکره‌های موازی[۲] تقسیم می‌شوند. پیکره‌های قیاس پذیر دارای متن معادلِ متن اصلی، به زبان‌های دیگر می‌باشند. در این نوع پیکره ها، متن معادل دقیقاً ترجمه‌ی متن اصلی نیست.

مجموعه‌ی ابهام برای هر واژه[۳] مجموعه‌ای است متشکل از واژه‌های صحیح فارسی که در فاصله ویرایشی مشخصی نسبت به واژه‌ی اصلی (به عنوان مثال فاصله‌ی یک) قرار دارند. تهیه مجموعه‌های ابهام اهمیت بسیاری به عنوان منابع زبانشناسی دارد زیرا در کاربردهایی مانند شناسایی نوری نویسه‌ها، تبدیل گفتار به متن، غلط یابی و تصحیح متن و ... کاربرد دارد (Dastgheib et al., 2016; Feili & Ghassem-Sani, 2004).

برای این منظور از خروجی قسمت اول این پژوهش یعنی لغت‌نامه استفاده می‌شود. مجموعه واژه‌های فارسی بدست آمده در این مرحله برای تولید مجموعه‌ی ابهام استفاده می‌شود. جدول ۳-۴ نمونه‌ای از مجموعه‌ی ابهام برای واژه‌ی شهد را نشان می‌دهد.

جدول ۳- ۴-مجموعه‌ی ابهام برای واژه‌ی «شهد»

واژه‌های مجموعه ابهام
شهد
شهر
عهد
مهد
شود

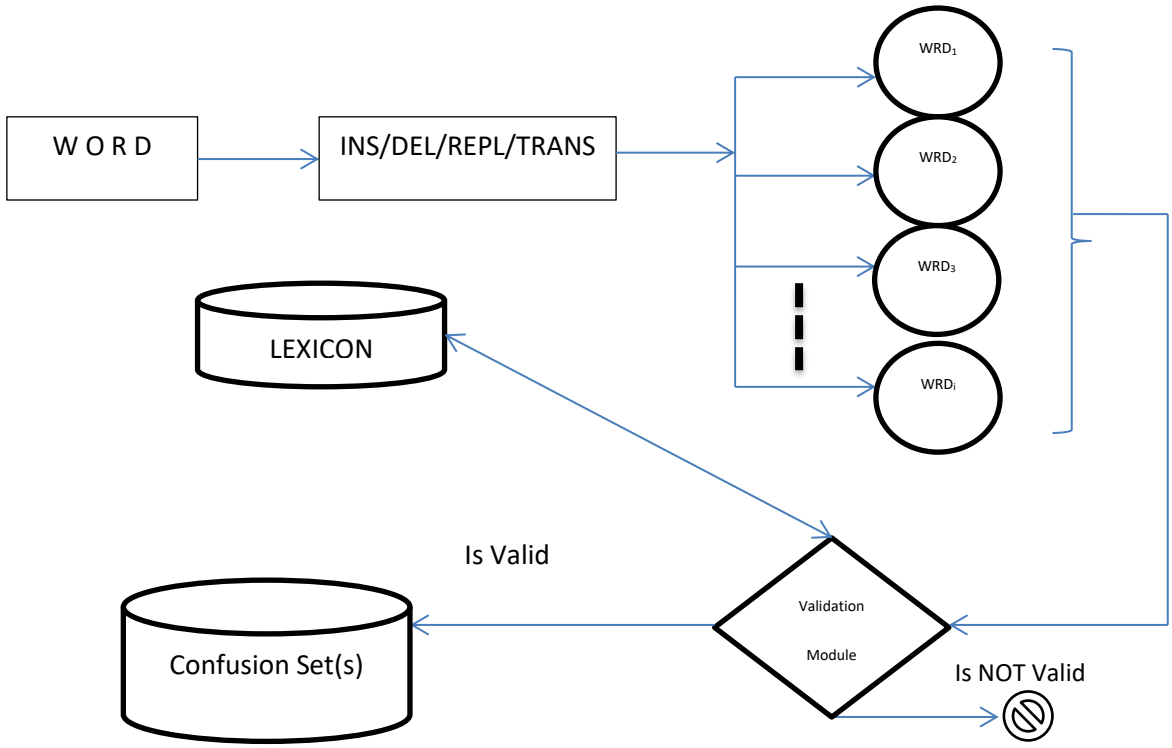
[۱] Comparable corpus
[۲] Parallel corpus
[۳] Confusion set

الگوریتم‌های متعددی برای تهیه واژه‌های مبهم بکار گرفته شده است که مرسوم ترین آن‌ها روش مستقیم و روش معکوس است. در روش مستقیم، فاصله‌ی واژه‌های لغتنامه با واژه‌ی هدف محاسبه می شود و اگر در فاصله‌ی ویرایشی مورد نظر (مثلا یک) باشد به مجموعه‌ی ابهام اضافه می‌شود. با توجه به اینکه بیش از ۸۵ درصد از خطاها در فاصله ویرایشی یک هستند (Dastgheib et al., 2016; Faili, Ehsan, Montazery, & Pilehvar, 2014; Gorin, 1971)، بنابراین معمولاً فاصله ویرایشی یک برای تولید پیکره استفاده می‌شود. مشکل روش مستقیم هزینه بسیار زیاد پردازش به ازای هر واژه است. هزینه تولید مجموعه ابهام برای هر واژه N بار جستجو در لغتنامه است و اگر برای کل واژه‌های لغت‌نامه بخواهیم مجموعه‌ی ابهام را تولید کنیم هزینه نهایی $O(N^2)$ خواهد بود. بنابراین این روش بخصوص اگر بخواهد بصورت برخط استفاده شود بسیار کند است و بهتر است پیکره‌ی مجموعه‌ی ابهام به عنوان یک منبع برون خطی محاسبه شود.

روش معکوس، با استفاده از واژه‌ی هدف مجموعه‌ی ابهام را تولید می‌کند(Gorin, 1971). در این روش همانطور که در شکل ۴-۳ نیز نشان داده شده است، با استفاده از چهار عملگر درج، حذف، جابجایی و جایگزینی برای حروف فارسی، واژه‌های جدید تولید کرده و اگر واژه‌ی صحیح باشد (کنترل با لغت‌نامه مرجع) به مجموعه‌ی ابهام اضافه می‌شود. در این روش اگر طول واژه L باشد، آنگاه ۴ اپراتور بروی هر حرف اعمال می‌شود و در نهایت L*۴ واژه با گنجینه لغت کنترل می‌شود که با توجه به نوع پیاده سازی گنجینه لغت در این پژوهش هزینه جستجو ناچیز است و در نهایت محاسبه امکان پذیر است.

روش معکوس سرعت بیشتر و هزینه‌ی کمتری دارد و نیاز به محاسبه برای کل لغت‌نامه ندارد. بنابراین در این پژوهش از این روش برای تولید مجموعه‌ی ابهام استفاده می‌شود. برای ذخیره مجموعه‌ی ابهام از ساختار

مجموعه‌ای در پایگاه داده‌ها استفاده می‌شود تا پس از ذخیره رکوردهای اطلاعاتی واژه‌ها، بسرعت و با کمترین هزینه قابل بازیابی باشد.



شکل ۴-۳-الگوریتم تولید مجموعه‌ی ابهام برای واژه‌های فارسی

همانطور که در شکل ۴-۳ نشان داده شده است، بازای هر نویسه‌ی واژه، چهار عملگر اعمال می‌شود و چندین رشته‌ی جدید تولید می‌شود، این رشته‌ها با کمک لغتنامه ارزیابی می‌شود و در صورتی که یک واژه‌ی معتبر باشد، در مجموعه‌ی ابهام ذخیره می‌شود.

فصل چهارم

نتایج

۴-۱-مقدمه

با توجه به لزوم وجود منابع و ابزارها مانند پیکره‌های موازی، پیکره‌های تک زبانه فارسی و ابزارهای پردازش زبان، که پیش‌نیاز تحقیق در حوزه‌ی زبان‌شناسی رایانه‌ای فارسی است، انجام پژوهش در راستای ساخت پیکره‌ها و ابزارهای پردازش زبان ضروری به نظر می‌رسد. در این راستا در این پژوهش، مجموعه‌ای از ابزار و پیکره فارسی تولید شده است.

برای مشخص شدن کارآیی و اثربخشی روش ترکیبی، یک دسته از آزمون ترتیب داده شده است. برای انجام این آزمون‌ها نیاز به یک مجموعه علامت‌گذاری شده برای اجرای آزمون‌ها است. این مجموعه بصورت علامت گذاری می‌گردد. برای این منظور مجموعه ای شامل ۱۵۰ جمله ترتیب داده شده است که با استفاده از آن بتوان تجریه‌کننده را ارزیابی نمود. مشخصات این مجموعه داده در جدول ۴-۱ نشان داده شده است.

جدول ۴-۱-مشخصات مجموعه داده‌ی آزمون تجزیه کننده

تعداد جمله	۱۵۰
تعداد واژه	۱۳۵۴
تعداد کاراکتر	۸۷۷۵
میانگین طول واژه	۴,۳

۴-۲-ارزیابی تجزیه کننده

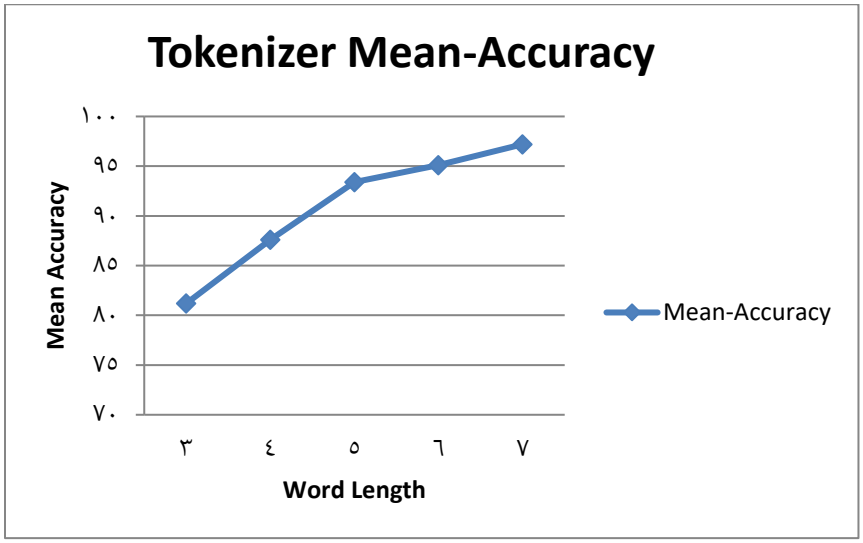
برای ارزیابی تجزیه کننده از مجموعه‌ی مشخص شده در جدول ۴-۱ استفاده شده است. برای ارزیابی تجزیه‌کننده هر جمله به صورت مستقل به تجزیه کننده داده شده است و نتیجه برای هر جمله ذخیره شده و در نهایت میانگین دقت تجزیه کننده برای این مجموعه بصورت نظارت انسانی محاسبه شده است. دقت تجزیه کننده برای هر واژه نیز ذخیره شده است. نهایتا دقت با استفاده از فرمول $P=TP/(TP+FP)$ محاسبه شده است. و درنهایت برای واژه‌ها و جملات میانگین دقت محاسبه شده است.

۴-۲-نتایج تجزیه کننده

واحد نتیجه	میانگین دقت
جمله	۹۵,۹
واژه ابتدای جمله	۹۹,۲
واژه انتهای جمله	۹۸,۶
واژه میان جمله	۸۹,۹

شکل ۴-۱ نیز دقت تجزیه‌کننده را بر اساس طول واژه نشان می‌دهد. همانطور که مشاهده می‌شود با افزایش طول واژه، دقت افزایش می‌یابد. با افزایش طول واژه، اطلاعات بیشتری در اختیار الگوریتم تجزیه کننده قرار می‌گیرد و بنابراین دقت افزایش می‌یابد. باید توجه داشت که فرکانس واژگان با طول کوتاه، در زبان فارسی بیشتر از واژگان طولانی است و طبق قانون زیف(Zipf, 1935) نیز فرکانس بکارگیری واژگان با طول کمتر نیز

بیشتر است. بنابراین افزایش دقت در این نوع واژه‌ها مهم است. لازم به ذکر است کارآیی ابزار Step-1 که در زمینه‌ی تجزیه‌کننده‌ی فارسی کاربرد دارد، ۰,۸۷ گزارش شده است (Shamsfard et al., 2010).



شکل ۴-۱-دقت تجزیه‌کننده براساس طول واژه

یکی از چالش‌های تجزیه‌کننده، تصحیح نویسه‌ی فاصله با نویسه‌ی نیم‌فاصله است. روش ترکیبی پیشنهادی دقت ۸۸,۴ را برای تصحیح نویسه فاصله با نیم‌فاصله کسب کرده است.

۴-۳-پیکره‌ی مجموعه‌ی ابهام واژه‌های فارسی

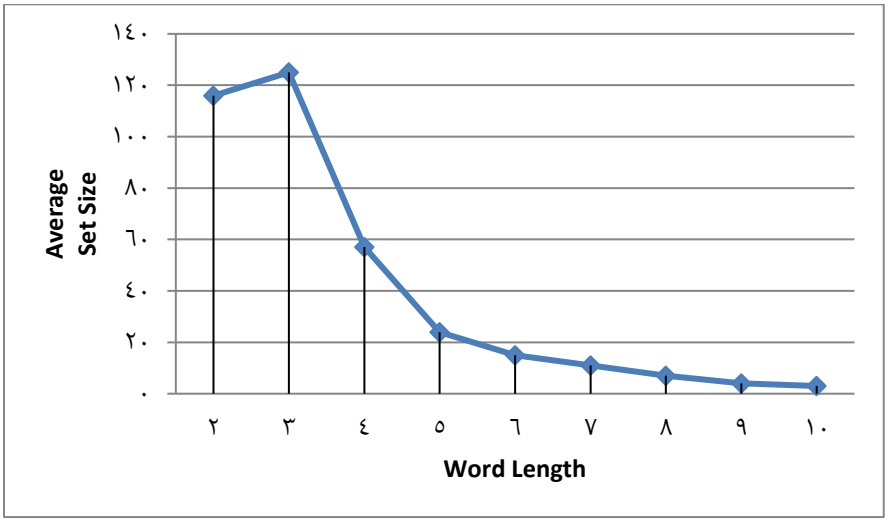
یکی از دستاوردهای مهم این پژوهش، پیکره‌ی مجموعه‌ی ابهام واژه‌های فارسی است. در این پیکره، به ازای هر واژه‌ی صحیح فارسی مجموعه‌ای از واژه‌های صحیح فارسی که در فاصله‌ی ویرایشی یک قرار دارند، ذخیره شده است. این مجموعه شامل بیش از ۱,۲ میلیون مجموعه‌ی ابهام در فاصله‌ی ویرایشی یک است. جدول ۴-۳ مشخصات پیکره را نشان می‌دهد.

جدول ۴-۳-مشخصات پیکره‌ی مجموعه‌ی ابهام واژه‌های فارسی

تعداد مجموعه‌ها	۱۲۰۰۰۰۰
میانگین تعداد عناصر در مجموعه‌های ابهام	۸
میانگین طول واژه‌ها	۶
میانگین طول واژه‌ها در مجموعه ابهام	۶
تعداد عضو کمترین مجموعه ابهام	۲
تعداد عضو بیشترین مجموعه ابهام	۴۵۸

همانطور که در شکل ۴-۲ نیز نشان داده شده است، برای واژه‌های بطول ۳، بیشترین فراوانی را دارد بنابراین مجموعه‌ی ابهام آن نیز دارای بیشترین اعضا در بین مجموعه‌های ابهام است. با افزایش طول واژه تعداد

واژه‌های مجموعه‌ی ابهام نیز کاهش می‌یابد زیرا با توجه به فراوانی، تعداد واژه‌هایی که در فاصله ویرایشی یک نسبت به واژه‌ی هدف هستند، کاهش می‌یابد.



شکل ۴-۲-میانگین تعداد عناصر مجموعه‌ی ابهام براساس طول واژه

در قسمت ضمیمه بخشی از پیکره که با قالب XML تهیه شده است، آورده شده است. XML مخفف زبان نشانه‌گذاری قابل گسترش می‌باشد. این زبان یک زبان نشانه‌گذاری جدید است که توسط کنسرسیوم وب در سال ۱۹۹۷ برای غلبه بر محدودیت‌های زبان HTML بوجود آمده است. کنسرسیوم وب سازمانی است که مسئول نگهداری استانداردهای موجود در زمینه وب می‌باشد که از مهمترین این استانداردها می‌توان به HTML اشاره کرد.

تفاوت اصلی XML با HTML در این است که زبان نشانه‌گذاری قابل گسترش سعی دارد داده‌ها را طوری نشانه‌گذاری کند، که معنای آن‌ها حفظ شود و در حالیکه HTML داده‌ها را طوری نشانه‌گذاری می‌کند که قابل نمایش برای مرورگرها باشد. در واقع تاکید زبان نشانه‌گذاری قابل گسترش بر روی معنای داده‌هاست در حالیکه

تاکید HTML بر نمایش داده می‌باشد. به منظور حفظ معنای داده‌ها زبان نشانه‌گذاری قابل گسترش ابرداده (Metadata) توصیف کننده داده‌ها را نیز همراه آن‌ها ذخیره می‌کند.

زبان نشانه‌گذاری قابل گسترش زیر مجموعه ساده شده‌ای از زبان SGML می‌باشد. SGML یک زبان عمومی و پیچیده برای نشانه‌گذاری داده‌هاست که در دهه ۸۰ بوجود آمد و پدر زبان‌های نشانه‌گذاری محسوب می‌شود. قابلیتهای زیاد این زبان آن را بیش از اندازه پیچیده کرده است بطوری که کنسرسیوم وب آن را بعنوان جانشین HTMLمناسب ندانسته و تصمیم گرفت زیر مجموعه ساده شده‌ای از آن را با نام XML جانشین HTML کند. زبان نشانه‌گذاری قابل گسترش، در واقع یک ابر زبان، نامیده میشود چرا که کاربر بسته به نیازهایی که دارد می‌تواند با استفاده از آن زبان نشانه‌گذاری جدیدی برای نشانه‌گذاری داده‌هایش ایجاد کند.

زبان نشانه‌گذاری قابل گسترش همچنین قادر است ساختار داده‌های ذخیره شده را نیز بهمراه معنای آن‌ها حفظ کند. این زبان دارای هیچ برچسب از پیش تعریف شده‌ای نیست و تمامی برچسب‌ها برحسب نیاز توسط کاربر تعریف می‌شوند. قابلیتهای زبان نشانه‌گذاری قابل گسترش و اجزاء همراه آن، این زبان را به زبان قابل حمل و استاندارد برای کاربردهای مختلف تبدیل کرده است.

فایل‌های زبان نشانه‌گذاری قابل گسترش دارای قابلیت متنی هستند طوری که می‌توان آنها را در ویرایشگرهای متنی ویرایش کرد. یک فایل در زبان نشانه‌گذاری قابل گسترش از دو قسمت متن و علائم نشانه گذاری تشکیل شده است که قسمت متن آن داده‌های اصلی ذخیره شده و علائم نشانه‌گذاری و ابرداده توصیف کننده متن را در بر دارد .زبان نشانه‌گذاری قابل گسترش از یک سو با ذخیره فایل‌هایش در قالب متنی و استفاده

از علائم نشانه‌گذاری فهم معنای داده‌های ذخیره شده را برای انسان ممکن می‌سازد و از سوی دیگر با در اختیار قرار دادن این فایل‌ها در یک قالب ساخت‌یافته برای برنامه‌ها، پردازش آن را برای کامپیوتر ساده می‌کند.

جزء اصلی تشکیل دهنده زبان نشانه‌گذاری قابل گسترش ، عنصر نام دارد که شامل نام و محتوی می‌باشد. محتوای یک عنصر بین دو علامت نشانه‌گذاری خاص با نام‌های برچسب شروع و برچسب پایان محصور می‌شود. روش برچسب‌گذاری زبان نشانه‌گذاری قابل گسترش همانند HTML است که در اصل HTML و زبان نشانه‌گذاری قابل گسترش این روش را از SGML به ارث برده‌اند. به دلایل ذکر شده در این پژوهش از ساختار زبان نشانه‌گذاری قابل گسترش برای پیکره استفاده شده است.

فصل پنجم

نتیجه گیری

۵-۱-نتیجه گیری

با توجه به لزوم وجود منابعی مانند پیکره‌های موازی، پیکره‌های تک زبانه فارسی و ابزارهای پردازش زبان، که پیش‌نیاز تحقیق در حوزه‌ی زبان‌شناسی رایانه‌ای فارسی است، انجام پژوهش در راستای ساخت پیکره‌ها و ابزارهای پردازش زبان ضروری به نظر می‌رسد (Dastgheib et al., 2016; Kashefi, Sharifi, & Minaie, 2012; Shamsfard et al., 2010).

ساخت ابراز تجزیه کننده، پیش‌نیاز بسیاری از پژوهش‌های پردازش زبان طبیعی است. ساخت پیکره‌هایی مانند مجموعه‌های ابهام در صورتی که به صورت تطبیق انسانی ساخته شود، بسیار پر هزینه است. از طرفی ساخت پیکره‌ها از متون موجود در وب مشکلات زیادی دارد که برخی از این مشکلات عبارت است از: متن غیر رسمی، درج و یا حذف شدگی زیاد (نویز)، کیفیت پایین متون تحت وب و نامعتبر بودن منابع.

در این پژوهش سعی شده است تعدادی ابزار و منبع که مورد نیاز پردازش زبان طبیعی برای فارس زبانان است، تهیه گردد. در این پژوهش یک پیکره‌ی ابهام واژه‌های فارسی، و ابزار تجزیه کننده تهیه شده است. برای تجزیه جملات به واژه و همچنین تهیه پیکره مجموعه‌ی ابهام از روش ترکیبی خودکار مبتنی بر دانش (پیکره) و مبتنی بر قانون استفاده شده است. این روش هزینه‌ی بسیار کمتری نسبت به روش تطبیق انسانی دارد و روش‌های دیگر مانند مبتنی بر قانون دارد و دقت مناسبی را کسب کرده است.

همانطور که در فصل چهارم ذکر شده است دقت نهایی حدود ۹۵ درصد برای واژه‌ها است که دقت مناسبی است. مجموعه‌ی ابهام نیز حدود ۱,۲ میلیون واژه‌ی صحیح فارسی را دربر دارد که با فرمت XML قابل استفاده توسط رایانه است.

با توجه به اینکه زبان فارسی دارای منابع زبانی الکترونیکی غنی برای پژوهش های زبان شناسی رایانه‌ای نیست، لذا پیشنهاد می‌شود در ادامه‌ی این پژوهش، ابزارهای دیگری مانند مصحح لغوی متن (این مصحح قاعدتاً باید بصورت تابع محور باشد تا به صورت جزء سوم نرم‌افزار قابل استفاده باشد)، ریشه‌یاب (ریشه‌یاب مورد نیاز پژوهش‌های زبان شناسی آماری است که بخش بزرگی از پژوهش‌های زبان شناسی رایانه‌ای را شامل می‌شوند) و پیکره‌هایی که شامل متون فارسی هستند (تولید پیکره‌های متنوع موجب افزایش منابع و دسترسی پژوهشگران به منابع با کیفیت خواهد شد) مانند پیکره برچسب دار موضوعی تهیه شود. مجموعه‌ی این پژوهش‌ها می‌تواند موجب افزایش کیفیت پژوهش‌های زبان‌شناسی رایانه‌ای با افزایش منابع گردد و راه حلی برای رفع نیاز به منابع و ابزارهای زبان فارسی است که پیش‌نیاز پژوهش در حوزه‌ی زبان شناسی رایانه‌ای است.

۵-۲-کاربردهای عملی پژوهش

همانطور که در فصل چهارم بیان شد، تولید پیکره و ابزار تجزیه کننده می تواند برای پیش پردازش متون فارسی بسیار مهم و حیاتی باشد. با توجه به اینکه اغلب پردازش های آماری در حوزه ی پردازش زبان طبیعی به روی واژه و فرکانس آن تمرکز دارد، لذا محاسبه دقیق فرکانس و بسامد واژه های فارسی امری مهم و اجتناب ناپذیر است. بنابراین، تقطیع جمله به واژه، که توسط این ابزار و با در نظر گرفتن نیم فاصله انجام می شود، می تواند در راستای این امر مهم نیز پژوهشگران در حوزه زبان شناسی رایانه ای را یاری نماید.

محصول دیگر این پژوهش، پیکره مجموعه ابهام واژه های فارسی است. این پیکره برای هر واژه فارسی، مجموعه ای از واژه هایی که در فاصله ویرایشی یک نسبت به واژه ی هدف قرار دارند را محاسبه نموده و در اختیار پژوهشگران قرار می دهد. این پیکره ماشین خوان، در غلط یابی متون ، شناسایی نوری نویسه ها، تصحیح غلط

های املایی، تبدیل گفتار به متن کاربرد دارد. این پیکره با قالب اکس ام ال، ارائه شده است که خود توصیف بوده و براحتی توسط زبان شناسان قابل استفاده خواهد بود. لازم به ذکر است، محاسبه واژه های مجموعه ابهام فرآیندی زمانبر است و بنابراین وجود این پیکره منحصربفرد که در فارسی نمونه آن وجود ندارد، منبعی ارزشمند است.

۵-۳-زمینه هایی برای مطالعه بیشتر

در آینده می توان این پژوهش حاضر را گسترش داد در این راستا، می توان برخی دیگر از خطاهای املایی که در این پژوهش به آن پرداخته نشده مانند مصوت های کوتاه پرداخت. همچنین می توان فاصله ویرایشی بیشتر از یک را نیز لحاظ نمود و پیکره حاضر را گسترش داد.

در راستای تولید ابزار نیز، زبان فارسی نیازمند ابزارهای بسیاری مانند ریشه یاب، غلط یاب و ... است. بنابراین می توان در راستای گسترش ابزارها نیز چنین ابزارهایی را به صورت تابع قابل فراخوانی در زبان های دیگر فراهم نمود و در راستای رفع کمبود ابزارهای فارسی قدم های مهمی برداشت. در این راه، در ادامه این طرح، تولید چنین ابزارهایی پیشنهاد می شود. بخصوص که ابزارهای مبتنی بر پیکره و روش های آماری، برای زبان فارسی می تواند به کارآیی و پاسخ مناسب برسد.

ضمیمه ۱- استفاده از تجزیه‌کننده

برای استفاده کننده یک DLL همراه با یک بانک اطلاعاتی واژه‌های فارسی که با فرمت جی‌سان[۱] ذخیره شده است، در اختیار پژوهشگران قرار گرفته است. بنابراین برای بکارگیری آن می‌توان در زبان‌های برنامه‌نویسی مختلف DLL را ضمیمه کرد و از فراخوانی تابع تجزیه کننده استفاده کرد. خروجی تابع تجزیه‌کننده، آرایه‌ای از رشته است. می‌توان واژه‌های تجزیه شده را به صورت لیستی از رشته‌ها دریافت کرد. در ذیل مثالی از بکارگیری تجزیه کننده و تولید خروجی XML آورده شده است. در ضمن با مراجعه به آدرس NLP.RICEST.AC.IR می توان از نسخه وبی استفاده کرد که خروجی XML را ارائه می‌نماید.

```
using System.Collections.Generic;
using System.Linq;
using System.Web;
using System.Web.Mvc;
using tokenizer;
using Newtonsoft.Json;
using System.IO;
using NLP.Models;
namespace NLP.Controllers
{

        public string tokenize(string s)
        {
            if (s == null) return ““;
            String G = “<XML>“;
            string DSTR=System.IO.File.ReadAllText(Server.MapPath(@”~/App_Data/LI.Txt”));
            tokenizer.NLP_TOOLS L= new NLP_TOOLS(DSTR);
            string [] WRDS=tokenizer.NLP_TOOLS.tokenize(s);
            G += “<SENTENCE>“;
            foreach(string j in WRDS)
            {
                G += “<WORD CONTENT='“+j+”'/>“;
            }
            G += “</SENTENCE>“;
            G += “</XML>“;
            return G;
        }

}
```

[۱] Json

منابع

Amtrup, J. W., Rad, H. M., Megerdoomian, K., & Zajac, R. (2000). Persian-English machine translation: An overview of the Shiraz project. *Memoranda in Computer and Cognitive Science MCCS-00-319, NMSU, CRL*.

Anvari, H. and Givi, H.A. (2006). Persian Language Grammar, Institute of Fatemi.

Asghar, M. Z., Khan, A., Ahmad, S., & Kundi, F. M. (2013). Preprocessing in natural language processing. *Editorial Board*, *152*.

Dastgheib, M. B., Fakhrahmad, S. M., & Jahromi, M. Z. (2016). Perspell: A new Persian semantic-based spelling correction system. Digital Scholarship in the Humanities, 32(3), 543-553.

Eghbalzadeh, H., Hosseini, B., Khadivi, S., & Khodabakhsh, A. (2012). Persica: A Persian corpus for multi-purpose text mining and Natural language processing. In *Telecommunications (IST), 2012 Sixth International Symposium on* (pp. 1207–1214).

Faili, H., Ehsan, N., Montazery, M., & Pilehvar, M. T. (2014). Vafa spell-checker for detecting spelling, grammatical, and real-word errors of Persian language. *Literary and Linguistic Computing*, fqu043.

Feili, H., & Ghassem-Sani, G. (2004). An application of lexicalized grammars in English-Persian translation. In *ECAI* (Vol. 16, p. 596).

Gorin, R. E. (1971). SPELL: A spelling checking and correction program. *Online Documentation for the DEC-10 Computer*.

Karimi, M., Tabrizi, H. H., & Chalak, A. (2016). Challenges in English to Persian Translation of Contracts and Agreements: The Case of Iranian English Translation Students. Journal of Applied Linguistics and Language Research, 3(6), 188-198.

Kashefi, Nasri, M., & Kanani, K. (2010). Towards Automatic Persian Spell Checking.

Tehran, Iran: SCICT.

Kashefi, O., Sharifi, M., & Minaie, B. (2012). A novel string distance metric for ranking Persian respelling suggestions. *Natural Language Engineering*, *19*(02), 259–284. doi:10.1017/S1351324912000186

Mahboubi, P., Compton, R. F., & Lu, T. C. (2017). System and method for Farsi language temporal tagger, U.S. Patent No. 9,740,689. Washington, DC: U.S. Patent and Trademark Office.

Megerdoomian, K. (2000). “Unification-Based Persian Morphology,” in CICLing.

Noaman, H. M., Sarhan, S. S., & Rashwan, M. (2016). Automatic arabic spelling errors detection and correction based on confusion matrix-noisy channel hybrid system. Egypt Comput Sci J, 40(2), 2016.

Mousavi, Z., & Faili, H. (2017). Persian Wordnet Construction using Supervised Learning. *arXiv Preprint arXiv:1704.03223*.

Rello, L., Romero, E., Rauschenberger, M., Ali, A., Williams, K., Bigham, J. P., & White, N. C. (2018, April). Screening dyslexia for English using HCI measures and machine learning. In Proceedings of the 2018 International Conference on Digital Health (pp. 80-84). ACM.

Sennrich, R., & Volk, M. (2010). MT-based sentence alignment for OCR-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, Colorado*.

Seraji, M., Megyesi, B., & Nivre, J. (2012). A basic language resource kit for Persian. In *Eight International Conference on Language Resources and Evaluation (LREC 2012), 23-25 May 2012, Istanbul, Turkey* (pp. 2245–2252).

Shamsfard, M. (2008). Developing FarsNet: A lexical ontology for Persian. In *4th Global WordNet Conference, Szeged, Hungary*.

Shamsfard, M. (2011). Challenges and open problems in Persian text processing. *Proceedings of LTC*, *11*.

Shamsfard, M., Jafari, H. S., & Ilbeygi, M. (2010). STeP-1: A Set of Fundamental Tools for Persian Text Processing. In *LREC*.

Smith, J. R., Quirk, C., & Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 403–411).

Taghva, K., Beckley, R., & Sadeh, M. (2005, April). A stemming algorithm for the Farsi language. In null (pp. 158-162). IEEE.

Taghva, K., Young, R., Coombs, J., Pereda, R., Beckley, R., & Sadeh, M. (2003, April). Farsi searching and display technologies. In Proc. of the 2003 Symp. on Document Image Understanding Technology (pp. 41-46).

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018, April). The Microsoft 2017 conversational speech recognition system. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5934-5938). IEEE.

Yannakoudakis, E. J., & Fawthrop, D. (1983). The rules of spelling errors. Information Processing & Management, 19(2), 87–99.

Zipf, G. K. (1935). The psycho-biology of language.

ستوده و هنرجویان، بررسی تنوع الگوهای نگارش فارسی و تاثیر آن بر جامعیت بازیابی اطلاعات، ۱۷(۲)، ۱۳۹۱.