استخراج اطلاعات از پیکره زبانی:
معرفی پیکرهٔ مقاله‌های علمی‌پژوهشی دانشگاه فردوسی مشهد

دکتر عطیه کامیابی گل[۱]، دکتر الهام اخلاقی باقوجری[۲]، دکتر احسان عسگریان[۳]، هانیه حبیبی[۴]

تاریخ ارسال:۹۶/۱۲/۷ ـ تاریخ پذیرش:۹۷/۲/۲۰

چکیده

هدف: پردازش زبان طبیعی، استفاده در فرهنگ‌نگاری، پیگیری تحولات زبانی و استخراج اطلاعات زبانی خاص از مهم‌ترین کاربردهای پیکره است. هدف از انجام این پژوهش معرفی و توصیف چگونگی ساخت پیکرهٔ مقاله‌های علمی‌پژوهشی است که نگارندگان پدید آورده‌اند.

روش: برای ایجاد پیکره، نخست نرم‌افزار پیکره‌ساز طراحی و ساخته شد. این نرم‌افزار انواع فرمت از جمله doc، docx، rtf، txt و pdf را پشتیبانی می‌کند. همچنین می‌توان پارامترهای ساخت پیکره را از قبل نیز برای آن تعیین کرد. برای مثال، مشخص کرد که حداقل تعداد توکن فایل برای حضور یک متن در پیکره چه عددی باشد. سپس مجموعه مقاله‌های علمی‌پژوهشی اعضای هیئت‌علمی دانشگاه فردوسی مشهد جمع‌آوری شد. مجموع مقاله‌های مشتمل بر ۷٫۱۵۴٫۲۰۲ کلمه و ۱۱۰۰ عنوان است. کلّ پیکره در فایل‌های مجزا به جمله‌های تشکیل‌دهنده آن تجزیه شد. ریشهٔ کلمات استخراج و برچسب نحوی کلمات زده شد. علاوه بر امکان استخراج اطلاعات به‌طور مستقیم، نرم‌افزار جانبی دیگری با کاربرد ساده‌تر برای استخراج اطلاعات آماری نیز طراحی و به آن افزوده شد تا کاربران غیرتخصصی هم بتوانند از آن استفاده و اطلاعات را استخراج کنند.

۱. استادیارگروه زبانشناسی دانشگاه فردوسی مشهد. kamyabigol@um.ac.ir
۲. دانش آموخته زبانشناسی دانشگاه فردوسی مشهد. elhamakhlaghi80@gmail.com
۳. دانش آموخته دکترای کامپیوتر دانشگاه فردوسی مشهد. ehsan.asgarian@gmail.com
۴. دانش آموخته مهندسی کامپیوتر و دانشجوی کارشناسی ارشد زبانشناسی دانشگاه فردوسی مشهد. hanieh.habibi@gmail.com

یافته: برای ارزیابی درستیِ ابزارهای ریشه‌یاب و برچسب‌زنی مقوله‌های گفتار، از پیکره‌های استاندارد موجود مانند پیکره PerDT (در سایت دادگان) که شامل تعداد قابل‌توجهی جملهٔ برچسب‌خورده با اطلاعات نحوی و ساخت‌واژی است استفاده شد. همچنین با مطالعه موردی عبارات احتیاط‌آمیز (بخشی از طرح پژوهشی که به چاپ نرسیده است) یافته این پژوهش که ساخت پیکرهٔ مقاله‌های علمی‌پژوهشی است آزمایش با دقت حدود۹۶٪ تأیید شد.

نتیجه: بر اساس نتایج به‌دست‌آمده، پیکرهٔ ساخته‌شده قابلیت بسیار بالایی برای داده‌کاوی و استفاده در تمام پژوهش‌هایی که بر روی متون علمی انجام می‌شود را دارا است. با استفاده از این پیکره می‌توان توصیفی داده‌محور از نحوهٔ کاربرد زبان توسط گروه‌های مختلف کاربران زبانی ارائه کرد. با بارگذاری این پیکره در سایت کتابخانه مرکزی دانشگاه فردوسی مشهد، امکان کاربری عام آن به‌زودی فراهم خواهد شد.

کلیدواژه‌ها: پیکرهٔ زبانی، برچسب‌دهی، مقاله‌های علمی‌پژوهشی، دانشگاه فردوسی مشهد.

۱. مقدمه

کمک به نیروی انسانی و کاهش خطا از دلایل مهمّ استفاده از رایانه است. امروزه ابزارهای سودمند و بسیار متداولی چون غلط‌یاب املایی و دستوری در ابتدایی‌ترین رایانه‌ها وجود دارد که موجب صرفه‌جویی در زمان و هزینهٔ کاربران می‌شود. اما همچنان میل به کم‌کوشی انسان را به سمت کاربردهای گسترده‌تر و تخصصی‌تر ابزارهای رایانه‌ای همچون بازشناسی خودکار متن[1] (تبدیل عکس به نوشتار) سوق داده است. کاربرد وسیع دیگری که درحال حاضر در حوزه تحقیق و پژوهش بسیار لازم و ضروری به‌نظر می‌رسد و مدتی است مورد توجه قرار گرفته است، امکان استفاده از انبوه داده‌های متنی است. با استفاده از این امکان، به‌جای استفادهٔ محدود از داده‌های متنی و یا صرف هزینه و زمان بسیار زیاد برای جمع‌آوری آن، با صرف زمان و هزینه اندک، از پیکره‌های موجود که دارای داده‌های انبوه و آمادهٔ استفاده و استخراج اطلاعات هستند، استفاده می‌شود. بدین‌ترتیب، محقق می‌تواند تمرکز اصلی خود را بر روی موضوعِ تحقیق قرار دهد و وقتش را صرف جمع‌آوری داده‌ها و اطلاعات نکند. بنا بر این ضرورت،

1. OCR

در این پژوهش که در حوزه زبانشناسی پیکره‌ای قرار می‌گیرد، پیکره ایجاد شده از مقاله‌های علمی‌پژوهشیِ اعضای هیئت‌علمی توصیف و معرفی می‌شود. منظور از پیکره، مجموعه‌ای نسبتاً بزرگ از متون الکترونیک است که به‌صورت حساب‌شده برچسب‌گذاری و دسته‌بندی شده‌اند و امکان بررسی‌های مختلف را به کاربر می‌دهند (مکنری ویلسون، ۲۰۰۱).

زبان‌شناسیِ پیکره‌ای سه حوزهٔ انجام پژوهش را پوشش می‌دهد که پژوهش حاضر دو حوزهٔ اول آن را شامل می‌شود:

ـ استخراج خودکار داده‌های زبانی از پیکره؛

ـ پردازش آماری داده‌ها؛

ـ ارزیابی و تفسیر این داده‌ها (تئوبری، ۱۹۹۱: ۱).

همچنین از منظر زبان‌شناسی، چهار حوزهٔ اصلی وجود دارد: آواشناسی، معناشناسی، ساخت‌واژه و نحو. پژوهش حاضر تمرکز خود را بر روی دو حوزهٔ ساخت‌واژه و نحو قرار داده است. گام نخست در حوزهٔ نحوی، شناسایی مقوله‌هایی است که واژه‌های یک زبان به آن تعلق دارند (آگرادی و دیگران، ۱۳۸۰: ۲۰۸). در حوزه ساخت‌واژه نیز نخستین گام، داشتن فهرستی از اقلام واژگانی است که مطلوب‌ترین حالت با استفاده از پایگاه‌های داده‌های زبانی عملی می‌گردد (عاصی، ۱۳۸۳).

برخی از مطرح‌ترین حوزه‌های پیکره‌ای در زبان فارسی عبارتند از: پیکرهٔ حوزهٔ اخبار (مانند پیکرهٔ همشهری)، پیکرهٔ ادبیات داستانی (مانند پیکرهٔ میزان)، پیکرهٔ رمان و زیرنویس فیلم، پیکره بیجن‌خان و... . هر کدام از این پیکره‌ها در حوزهٔ خاصی توسعه یافته‌اند. تاکنون پیکره مقاله‌های علمی‌پژوهشی پدید نیامده و همین عامل وجه تمایز پیکرهٔ حاضر با پیکره‌های موجود است. استفاده از پیکره حوزه‌های گسترده‌ای را شامل می‌شود و پژوهشگران مختلف براساس نیاز خود می‌توانند از پیکره استفاده کنند. به‌عنوان نمونه، «مشهور و فقیری» (۱۳۹۱) برای بررسی انواع زمان از پیکره استفاده کرده‌اند. موارد مورد استفاده از این پیکره نیز می‌تواند متناسب با نیاز کاربر، متنوع باشد.

۲. پیشینه

در چند دههٔ اخیر رویکردهای پیکره‌محور در زبان‌شناسی کاربردی استقبال شده‌اند، زیرا امکانات تحلیلیِ بسیار دقیقی برای زبان فراهم می‌کنند به‌طوری که بسیاری از کشورها اقدام به تهیهٔ پیکرهٔ زبان بومی خود کرده‌اند. علاوه بر این، بسیاری از پیکره‌های خردتر که اهداف پژوهشی خاصی را دنبال می‌کنند نیز به‌وجود آمده‌اند. پیکره‌های زبان‌آموز، موازیِ دوزبانه، متون ادبی و متون ترجمه‌شده از این دست هستند (گرینجر، گیلکوین و مونیر[1]، ۲۰۱۵). برای پیکره کاربردهای مختلفی را برشمرده‌اند. با استفاده از این نوع پیکره می‌توان استفادهٔ دقیق زبان‌آموز از نوع و میزان واژگان و ساخت‌های دستوری را مشخص کرد (همان). «بیکر» (۲۰۰۶) در کتاب خود کاربرد پیکره را در تحلیل گفتمان بررسی و پلی میان تحلیل گفتمان و زبان‌شناسی پیکره‌ای برقرار کرده‌است. یکی از کاربردهای دیگر پیکره، آموزش زبان است.

در ساخت پیکره مسائل مختلفی مطرح می‌شود؛ از جمله، دسته‌بندی مقولات گفتار مانند فعل. چگونگی نوع دسته‌بندی افعال و سایر مقوله‌های زبانی در تحقیقات مختلفی ارائه شده است. «فرخ» (۱۳۸۱)، دسته‌بندی مبسوطی از افعال زبان فارسی برای تشخیص خودکار رایانه ارائه کرده است. «آراسته» (۱۳۸۱) نرم‌افزار تشخیص فعل را طراحی کرده است. «حاجی و عبدالحسینی» (۲۰۰۰) از روش ریاضی و آمار برای دسته‌بندی مقوله‌های گفتار بهره برده‌اند. «رجا و همکاران» (۲۰۰۷) نیز برچسب‌گذاری متون زبان فارسی را بررسی کرده‌اند.

اما زبان فارسی به‌دلیل پیچیدگی‌های زبانی، کمبود منابع و مطالعات انجام‌شده از دیدگاه محاسباتی کمتر مورد توجه پژوهشگران قرار گرفته است و در آثار بسیار اندکی از جمله «شمس‌فرد» (۲۰۱۱) و «فیلی، منشادی، فردرکینگ[2] و لوین[3]»(۲۰۱۴) شاهد آن

1. Granger, Gilquin, & Meunier
2. Frederking
3. Levin

هستیم. متأسفانه ابزارهای استانداردِ پیش‌پردازشِ ایجادشده برای متون زبان فارسی از قبیل «شمس‌فرد، جعفری و ایل‌بیگی» (۲۰۱۴)، «سرابی، مهیار و فرهودی» (۲۰۱۳) و «سراجی، مقیسی[1] و نیور»[2] (۲۰۱۲) به‌صورت رایگان منتشرنشده‌اند. برخی از ابزارهای کد باز پیش‌پردازش موجود از قبیل «خلش و ایمانی» (۲۰۱۴) و «جدیدنژاد، محمودی و دهداری» (۲۰۱۰) نیز دقت مناسب را ندارد. فهرست پیکره‌های موجود در زبان فارسی را می‌توان در سایت http://dadegan.ir/catalog?page=3 مشاهده کرد هرچند دسترسی به همهٔ آنها امکان‌پذیر نیست.

۳. پیکره

وجود پیکره‌های زبانی در امر پردازش زبان یکی از ضروریات است. برخی از پیکره‌های موجود در این زمینه در ادامه فهرست شده است؛ البته هریک از این پیکره‌ها متفاوت است.

۱. پیکرهٔ حوزه خبر مرکز تحقیقات مخابرات ایران: حجم این پیکره ده میلیون لغت است و در آن متون ترازشدهٔ زبان فارسی در مقابل اخبار آورده شده است.

۲. پیکرهٔ میزان (ادبیات داستانی) مربوط به شورای عالی اطلاع‌رسانی: این پیکره پانزده میلیون لغت دارد و دوزبانه است.

۳. پیکرهٔ رمان و زیرنویس فیلم.

۴. پیکرهٔ بیجن‌خان: پیکره‌ای برچسب‌گذاری‌شده است که در حوزه پردازش زبان طبیعی به‌کار می‌رود و متشکل از ۴۳۰۰ موضوع مختلف از اخبار و متون عامیانه است. (پردازش زبان‌های طبیعی عبارت است از، استفاده از رایانه به‌منظور پردازش زبان گفتاری و نوشتاری).

۵. پیکره همشهری

---

1. Megyesi
2. Nivre

۶. بانک‌های درختی: پیکرهٔ درختی مجموعه‌ای از جمله‌های فارسی است که در آنها روابط نحوی کلمات بر مبنای نقش دستوری آنها مشخص شده است.

برخی پیکره‌های موجود هم در حوزهٔ صوت است (یاری، ۱۳۹۴).

باید خاطرنشان کرد، جمع‌آوری داده‌ها در یک مجموعه به خودی خود ارزش بسیار بالایی ندارد. ارزش این داده‌ها زمانی به حداکثر خود می‌رسد که اطلاعات جانبی یا مشخصه‌های زبانی مختلف به آن داده‌ها اضافه شده باشد که در اصطلاح به آن «حاشیه‌نویسی» گفته می‌شود. حاشیه‌نویسی معنایی با به‌کارگیری پردازش زبان طبیعی، یادگیری ماشین و یادگیری آماری داده‌های فاقد ساختار را که در قالب‌های مختلف مثل متن، تصویر، صوت و... منتشر شده‌اند، با افزودن فراداده‌ها غنی می‌کند. حاشیه‌نویسی در فرهنگ لغت آکسفورد، یادداشت توضیحی یا نظری‌ای که به یک متن یا نمودار اضافه شده، تعریف شده است. در یک برنامهٔ نرم‌افزاری حاشیه‌نویسی توضیحاتی است که به برنامه اضافه شده است. در یک عکس تبلیغاتی شعاری است که در ذیل آن اضافه شده و در یک نمودار اطلاعات تکمیلی است که به آن ضمیمه گردیده است. بر اساس روش انجام، می‌توان پلت‌فرم‌های حاشیه‌نگاری را به دو دستهٔ اصلی زیر تقسیم کرد:

ـ مبتنی بر الگو[1]: الگو می‌تواند کشف یا به‌طور دستی تعریف شود. بیشتر این روش‌ها از متدی که برین[2] معرفی کرده است، استفاده می‌کنند.

ـ مبتنی بر یادگیری ماشین[3] (متیو[4]، ۲۰۰۵).

همچنین با توجه به نحوه انجام کار، حاشیه‌نویسی معنایی را می‌توان به سه نوع دستی (در این روش عملیات حاشیه‌نویسی توسط نیروی انسانی انجام می‌گیرد)، نیمه‌خودکار (روش نیمه‌خودکار از نیروی انسانی در بخش‌هایی از فرایند انجام

1. Pattern Based
2. Brin
3. Machine Learning Based
4. Matthew

حاشیه‌نویسی بهره می‌برد) و خودکار (حاشیه‌نویسی بدون دخالت و نظارت نیروی انسانی را خودکار می‌نامند) تقسیم کرد (ازبک[1]، ۲۰۱۵؛ تالانتیکر[2]، ۲۰۰۹). به‌علاوه، حاشیه‌نویسی را می‌توان براساس محتوای منابعی که قرار است حاشیه‌نویسی شود نیز دسته‌بندی کرد (یون[3]، ۲۰۰۸؛ کیریاکف[4]، ۲۰۰۴). از این بُعد می‌توان متن، تصویر و تصاویر متحرک را که تاکنون بیشتر بر روی آنها فعالیت صورت گرفته است، نام برد. یکی از چالش‌های بزرگ در پردازش خودکار متن‌های زبانی، شناسایی واژه‌ها و نشانه‌گذاری آنهاست. نشانه‌گذاری دستوری را معمولا برچسب‌دهی می‌نامند (عاصی، ۱۳۸۳).

۱. پیکرهٔ مقاله‌های علمی

با درنظرگرفتن اهمیت موضوع، ابزار ساخت پیکرهٔ مقاله‌های هیئت علمی به‌وجود آمد. سپس با استفاده از آن، پیکره‌ای با تعداد ۱۱۰۰ مقاله از اعضای هیئت علمی دانشگاه فردوسی مشهد ساخته شد. فرایند ساخت این پیکره در ادامه بیان شده است.

۴-۱ روش جمع‌آوری داده‌ها

برای دستیابی به اهداف پژوهش، متون علمی اعضای هیئت علمی دانشگاه فردوسی مشهد (شامل ۱۱۰۰ مقاله) گردآوری و براساس رشته و دانشکده به دو دستهٔ کلی تقسیم شد. دستهٔ اول مقاله‌های حوزه‌های کشاورزی، دام‌پزشکی و علوم پایه و دسته دوم مقاله‌های حوزه‌های علوم انسانی و مهندسی را شامل می‌شود. هرچند امکان بررسی کلیِ همهٔ حوزه‌ها با هم وجود دارد، این تقسیم‌بندی برای رویکردهای مقایسه‌ای امکان مناسبی را فراهم کرده است. سپس با استفاده از نرم‌افزار پیکره‌سازِ طراحی‌شده، داده‌ها از

1. Usbeck
2. Talantikitr
3. Yun
4. Kiryakov

نظر مقوله‌های گفتار از جمله اسم، فعل، صفت، حرف اضافه و قید برچسب‌گذاری شدند. در نهایت، دو خروجیِ اصلیِ برچسب‌خورده به‌دست آمد که امکان جستجوی سریع تمام مطالب در آن وجود داشت.

۴-۲ ابزارهای پردازش متن فارسی

براساس تقسیم‌بندی‌های ذکرشده، ابزارهای حاشیه‌نویسیِ مبتنی بر متن متفاوتی به‌وجود می‌آید. استانداردسازی، تفکیک متن به جملات، عبارات و کلمات و برچسب‌گذاری و حاشیه‌نویسی آنها، تأثیر بسزایی بر پردازش و استخراج اطلاعات، دسته‌بندی یا دیگر کاربردهای پردازش زبان طبیعی دارد. بیش از صدمیلیون نفر از مردم جهان به زبان فارسی صحبت می‌کنند. فارسی زبان رسمی سه کشور ایران، افغانستان (فارسی دری) و تاجیکستان (فارسی تاجیکی) است. به‌دلیل پیچیدگی‌های زبانی، منابع و مطالعات انجام‌شده در این زبان از دیدگاه محاسباتی کمتر مورد توجه پژوهشگران قرار گرفته است (شمس‌فرد، ۲۰۱۱؛ فیلی، ۲۰۱۴).

متأسفانه ابزارهای استاندارد پیش‌پردازشِ ایجادشده برای متون زبان فارسی از قبیل شمس‌فرد (۲۰۱۰)، سرابی، مهیار و فرهودی (۲۰۱۳)، سراجی، مقیسی[1] و نیور[2] (۲۰۱۲) رایگان منتشر نشده‌اند. برخی از ابزارهای کد باز[3] پیش‌پردازش موجود از قبیل خلش و ایمانی (۲۰۱۴)، جدیدنژاد، محمودی و دهداری (۲۰۱۰) و منشادی (۲۰۱۵) نیز دقت مناسب را ندارند. در این بخش توضیحات ابزارهای تولیدشده و مورد نیاز برای نظرکاوی در زبان فارسی را ارائه می‌کنیم.

1. Megyesi
2. Nivre
3. Open Source

۳-۴ نرمال‌سازی و جداسازی جملات و کلمات متن

قبل از پردازش متون جهت استانداردسازی حروف و فاصله‌ها باید پیش‌پردازش‌هایی روی آنها انجام شود. در پردازش رسم‌الخط زبان فارسی، با توجه به قرابتی که با رسم‌الخط عربی دارد، همواره در نگارش تعدادی از حرف‌ها مشکل کاراکترهای عربی معادل وجود دارد. از جملهٔ آنها می‌توان به حروف «ک»، «ی»، همزه و... اشاره کرد. در گام نخست باید مشکلات مربوط به این حروف را با یکسان‌سازی آنها برطرف کرد. در این مرحله باید همهٔ نویسه‌ها (حروف) متن با معادل استاندارد آن جایگزین و یکسان‌سازی شود. علاوه بر این، اصلاح نویسهٔ نیم‌فاصله و فاصله در کاربردهای مختلف آن و همچنین حذف نویسهٔ «ـ» که برای کشش‌نویسه‌های چسبان استفاده می‌شود و یکسان‌سازی متون برای تشدید، تنوین و موارد مشابه (مشابه ابزار PrePer) (سراجی، ۲۰۱۰) از جمله اقدام‌های لازم قبل از شروع پردازش متن است.

در این فاز مطابق با یک‌سری قاعده دقیق و مشخص، فاصله‌ها و نیم‌فاصله‌های موجود در متن برای علاماتی نظیر "ها"، "تر" و "ی" غیرچسبان در انتهای لغات و همچنین پیشوندها و پسوندهای فعل‌ساز نظیر "نمی"، "می"، "ام"، "ایم"، "اید" و موارد مشابه نیز اصلاح می‌گردند. پس از پایان مرحله‌ی پیش پردازش متون، ابزار تشخیص‌دهندهٔ جمله‌ها با استفاده از علامت‌های "."، "؛"، "!"، "؟"، "?"، ":" و به‌کارگیری برخی قواعد دستوری زبان فارسی و در نظرگرفتن برخی واژگان آغازکنندهٔ جمله‌ها (از قبیل حروف ربط مانند "که"، "اساسا"، "البته"، "تا"، "اما"، "اگر"، "ولی"، "زیرا"، "سپس"، "همچنین"، "و" و "یا")، مرز جمله‌ها را تعیین می‌کند. تشخیص‌دهندهٔ واژگان نیز با استفاده از علامت‌های فضای خالی، "،"، ","، "-" و ... و درنظرگرفتن اصلاحات اعمال‌شده دربارهٔ پیشوندها و پسوندها در فاز قبلی، واژه‌ها را شناسایی می‌کند. همچنین پردازش ویژه‌ای برای درنظرگرفتن یک علامت برای کلمات اختصاری (از قبیل A.T.R یا بی.بی.سی)، تاریخ و زمان (از قبیل 5:35 یا 2015/2/25)، اعداد اعشاری (از قبیل 5/17 یا 5.17) و سایر عبارت‌ها و علایم خاص (جایگزینی کلمه "ا..." با کلمه اصلی آن) انجام می‌شود.

۴-۴ ریشه‌یابی

ریشه‌یابی واژگان از عملیات مهمّ پیش‌پردازش متون در بازیابی اطلاعات و پردازش زبان‌های طبیعی است. هدف الگوریتم‌های ریشه‌یابی، حذف پیشوند و پسوندهای کلمات و تعیین ریشهٔ اصلی کلمه است. توضیحات بیشتر دربارهٔ اهمیت ریشه‌یابی و تحلیل‌های ریخت‌شناسی کلمات در استخراج دانش از متن (عبدالمجید، دیاب[1] و کوبلر[2]، ۲۰۱۴) وجود دارد. ریشه‌یابی به‌طور گسترده در سیستم‌های بازیابی اطلاعات، ترجمه ماشین، دسته‌بندی متن، خلاصه‌نویسی متن، شاخص‌گذاری، متن‌کاوی و... استفاده می‌شود. برخلاف زبان انگلیسی، مشکلات مختلفی هنگام ریشه‌یابی کلمات زبان فارسی وجود دارد؛ از جمله اینکه ضمایر می‌توانند به دو صورت جدا و متصل در جمله ظاهر شوند. البته در مورد افعال مسئله کمی پیچیده‌تر است، بطوری که علاوه بر وندهای فعلی، شخص (فاعل) و زمان جمله نیز بر روی حالت فعل تأثیرگذار است.

در روش‌های ریشه‌یابی فعلی در زبان فارسی، پس از حذف وندها ممکن است معنای کلمه تغییر کند. در ریشه‌یاب (بن‌واژه‌یاب[3]) تولیدشده، ریشه‌یابی واژه بدون تغییر مفهوم واژه در جمله مبنای نظر است. همچنین الگوریتم بیان ارائه‌شده قابلیت تعیین ریشه در چند سطح را دارد. این سطوح مختلف ریشه می‌توانند در عملیات مختلف پردازش زبان طبیعی مورد استفاده قرار گیرند. در ابزار ریشه‌یاب تولیدشده، از دو رویکرد استفاده از فرهنگ لغت و قوانین ریخت‌شناسی و برای این ابزار، از پنج فرهنگ لغت استفاده شده است:

۱. فرهنگ لغت برای نگه داشتن تمامی ریشه های مربوط به کلمات غیرفعلی (شامل اسامی، صفت‌ها، قیدها).

۲. فرهنگ لغت برای نگهداری کلمات جمع مکسر و حالت‌های جمع بدون قاعده

---

1. Diab
2. Kubler
3. Lemmatizer

به همراه حالت مفرد این کلمات.

۳. فرهنگ لغت شامل ریشه‌های گذشته و حال افعال زبان فارسی.

۴. فهرست انواع وند (پیشوند و پسوند) و قوانین مربوط به ترتیب قرارگیری آنها در زبان فارسی.

۵. فرهنگ لغت برای کلمات استثنا (وندهای متعلق به ریشه یا هستهٔ اصلیِ کلمه).

در الگوریتم ریشه‌یابی، ابتدا باید نوع عبارت از نظر فعل یا غیرفعل تشخیص داده شود. برای شناسایی صحیح انواع فعل (پیشوندی، ساده و مرکب) در زمان‌ها و شکل‌های مختلف در متن، قواعد دستوری زبان فارسی، موقعیت لغات (بافت جمله) و تحلیل ریخت‌شناسی واژگان (مطالعه ساختار لغات) بررسی شد. بدین‌منظور، از فرهنگ لغت مربوط به بُن‌های گذشته و حال افعال زبان فارسی مجموعه «دادگان» (رسولی و همکاران، ۲۰۱۱) که حدود ۶۰۰۰ فعل ساده، پیشوندی و مرکب را شامل می‌شود، استفاده شده است.

برای ریشه‌یابی واژگان غیرفعل، ابتدا کلّ واژه درون فرهنگ لغت شامل ریشهٔ واژگان جستجو می‌شود. اگر واژه وجود داشته باشد، خود واژه به‌عنوان ریشه معرفی می‌شود. در غیر این‌صورت، واژه درون فرهنگ لغت مربوط به واژگان جمع مکسر و بدون قاعده جستجو می‌شود. اگر واژه در این فرهنگ لغت یافت شود، معادل مفرد آن به همراه یک نشانگر مبنی بر جمع مکسربودن آن، برگردانده می‌شود. سپس با استفاده از قوانین ریخت‌شناسی مربوط، تمامی وندهای ممکن موجود در واژه یافت می‌شود. در هر مرحله، پس از حذف پسوند، دوباره واژه درون فرهنگ لغت مربوط به ریشهٔ واژگان غیرفعل زبان فارسی جستجو می‌شود. چنانچه واژه یافت شود، می‌تواند به‌عنوان ریشه معرفی شود. همین مراحل برای پیشوندها نیز صورت می‌گیرد. پیشوندها نیز از واژهٔ مربوط حذف می‌شوند تا ریشه به‌دست آید. همچنین، ممکن است دو واژه در متن به هم متصل شده باشند و برنامه نتواند طبق روال فوق آنها را ریشه‌یابی کند. به‌منظور رفع این مشکل، برای جداسازی واژگان به‌هم‌چسبیده در متن، هر واژه که در فرهنگ ریشه واژگان یافت نشود،

به دو بخش بزرگ‌تر از دو حرف شکسته شده و درصورتی‌که هر دو بخش در فرهنگ ریشه واژگان وجود داشته باشد، عمل جداسازی انجام می‌شود. این عملیات به‌صورت تکراری ادامه می‌یابد تا در نهایت ریشهٔ نهایی استخراج گردد. با تکرار این عملیات، ابزار ریشه‌یابی این قابلیت را پیدا می‌کند که بتواند سطوح (ریشه‌ها) معانی مختلف واژه را (برای استفاده در کاربردهای گوناگون) استخراج کند. به‌عنوان مثال، برای واژهٔ «خلاصه‌سازی» سه ریشه «خلاصه‌سازی»، «خلاصه‌ساز» و «خلاصه» به‌عنوان خروجی بازگردانده می‌شود که بسته به نوع استفاده می‌توان از هریک از آنها استفاده کرد. به‌عنوان مثال دیگر، برای کلمه «دانشجویان» نیز سه ریشه «دانشجو»، «دانش» و «دان» استخراج می‌شود.

۴-۵ برچسب‌زنی اجزای کلام

برچسب‌زنی نقش اجزای کلام[1] عمل انتساب برچسب‌های نحوی به واژه‌ها و نشانه‌های تشکیل‌دهندهٔ یک متن است؛ به‌صورتی که این برچسب‌ها نشان‌دهندهٔ نقش واژگان و نشانه‌ها در جمله باشد. در زبان فارسی اغلب واژگان (حدود۹۱٪) در پیکرهٔ بیجن‌خان (۲۰۰۴) دارای نقشی واحد در جمله‌های مختلف هستند. سایر واژگان از نظر برچسب‌زنِ نحوی دارای ابهام هستند، زیرا ممکن است واژگان در جایگاه‌های مختلف برچسب‌های نحوی متفاوتی داشته باشند. بنابراین، برچسب‌زنی نحوی، عمل ابهام‌زدایی از برچسب‌ها با توجه به زمینه (ساختار جمله) مورد نظر است.

برای بررسی آمار تعداد تکرار واژگان و برچسب‌های مختلف به‌صورت مجزا یا در کنار هم و استخراج قوانین نحوی و ریخت‌شناسی (ساختار واژگان و ارتباط واژگان در جمله) در متون زبان فارسی از پیکرهٔ استاندارد دادگان[2] (رسولی، ۲۰۱۵) استفاده شد. همچنین فهرستی از انواع وندهای صفت‌ساز، قیدساز و اسم‌ساز در واژگان زبان فارسی برای

1. Part of Speech tagging
2. Dadegan Treebank

استفاده جهت پیش‌بینی احتمالات اولیه به واژگان جدید (خارج از پیکره)، شناسایی شد. برای شناسایی بهتر نقش واژگان جدید در جمله‌های عملیات پیش‌پردازش شامل نرمال‌سازی و استانداردسازی حروف و فاصله‌ها و ریشه‌یابی کلمات[1] روی مستندات پیکره انجام می‌شود.

ابزار تهیه‌شده برای برچسب‌گذاری نقش ادات سخن در متون فارسی، از پیکره برچسب‌خوردهٔ دادگان و از ترکیب دو روش مدل مخفی مارکوف و برچسب‌گذاری مبتنی بر قانون، استفاده می‌کند. برای برچسب‌گذاری اجزای کلام، ابتدا متن ورودی پیش‌پردازش می‌شود. سپس به کمک روش یادگیر HMM[2] (پیاده‌سازی‌شده به روش ویتربی) و با استفاده از اطلاعات آماریِ محاسبه‌شده از پیکره‌های برچسب‌خورده، محتمل‌ترین برچسب نقش ادات سخن به هر کلمه در جمله انتساب می‌یابد (Ganchev, Taskar, Pereira, and Gama, ۲۰۰۹). در مرحله بعد، این برچسب‌ها به‌وسیلهٔ دو گروه از قوانین نحوی و ریخت‌شناسی از پیش استخراج‌شده، اصلاح‌شده و برچسب نهایی نقش هر کلمه مشخص می‌شود. گروه اول قوانینی هستند که به‌صورت خودکار از پیکرهٔ برچسب‌خورده استخراج شدند (براساس روش برچسب‌زنی بریل (Megyesi، ۱۹۹۹) و گروه دوم شامل تعداد محدودی از قوانین نحوی است که زبان‌شناسان زبان فارسی آنها را استخراج کرده‌اند.

در مرحله بعد، این برچسب‌ها به‌وسیله دو گروه از قوانین نحوی و ریخت‌شناسی از پیش استخراج‌شده، اصلاح و برچسب نهایی نقش هر واژه مشخص می‌شود. گروه اول قوانینی هستند که به‌صورت خودکار از پیکرهٔ برچسب‌خورده استخراج شدند (براساس روش برچسب‌زنی بریل[3] (مقیسی، ۱۹۹۹). گروه دوم شامل تعداد محدودی از قوانین نحوی هستند که زبان‌شناسان زبان فارسی آنها را استخراج کرده‌اند. در شکل۱، نرم‌افزار

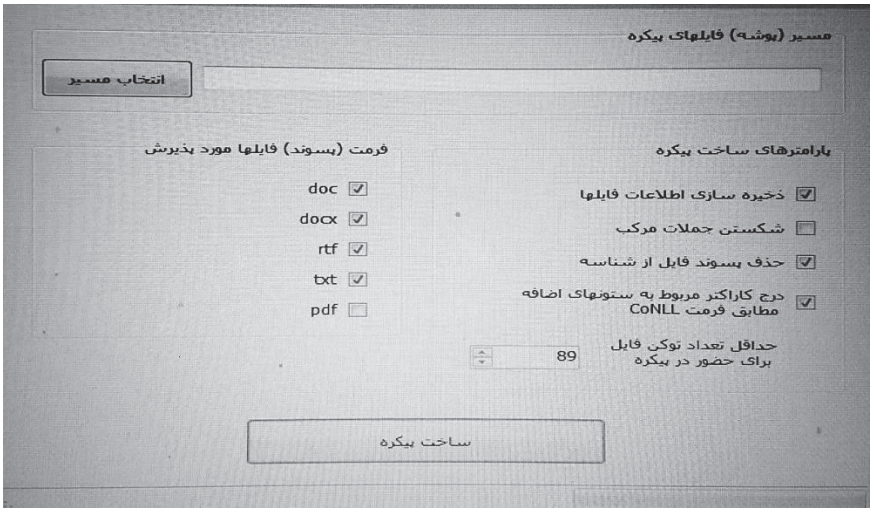
---

۱. برای تعیین احتمال اولیه نقش کلمات خارج از پیکره آموزشی عمل ریشه‌یابی نیز انجام می‌شود.

2. Hidden Markov Model

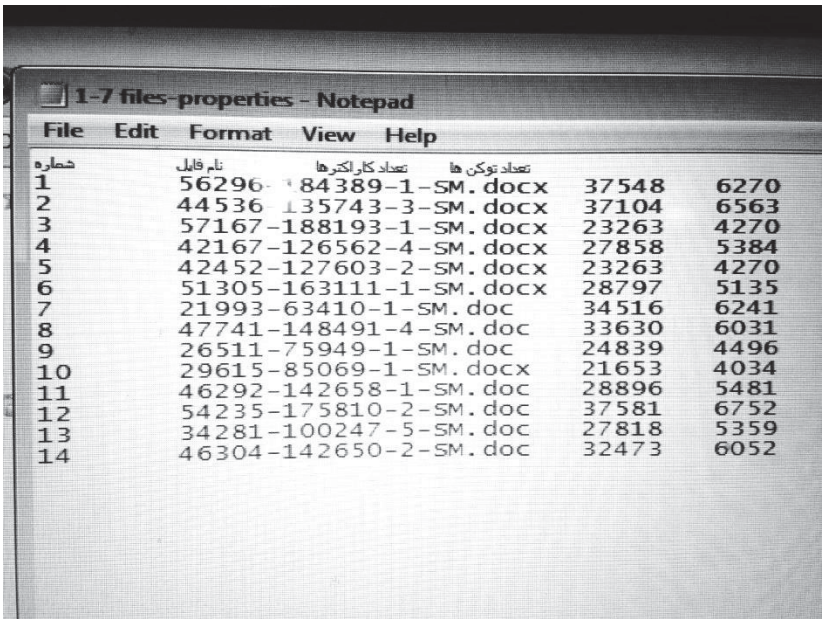
3. Brill POS-Tagger

تهیه‌شده برای ساخت پیکره را مشاهده می‌کنید.



شکل ۱. نمایی از نرم‌افزار ساخت پیکره

برای استفاده و استخراج اطلاعات از پیکره، یک نرم‌افزار جانبی طراحی شده است که در شکل ۴ نمایی از آن را مشاهده می‌کنید. در این امکان می‌توان با هم‌آیی، انواع ترکیب واژگان و سایر اطلاعات مورد نیاز را استخراج کرد. اما خود نرم‌افزار هم به دو شیوه خروجی اطلاعات را نشان می‌دهد که در شکل ۲ فرمت کلی آن را که نمایش عناصر موجود در یک فایل است، نشان داده شده است. فرمت دیگر آن تجزیهٔ اجزا به توکن‌ها، ریشه و روابط معنایی را بیرون می‌دهد.


1-7 files-properties - Notepad

File Edit Format View Help

شماره	نام فایل	تعداد کاراکترها	تعداد توکن ها	
1	56296	84389-1-SM.docx	37548	6270
2	44536	35743-3-SM.docx	37104	6563
3	57167-188193-1-SM.docx		23263	4270
4	42167-126562-4-SM.docx		27858	5384
5	42452-127603-2-SM.docx		23263	4270
6	51305-163111-1-SM.docx		28797	5135
7	21993-63410-1-SM.doc		34516	6241
8	47741-148491-4-SM.doc		33630	6031
9	26511-75949-1-SM.doc		24839	4496
10	29615-85069-1-SM.docx		21653	4034
11	46292-142658-1-SM.doc		28896	5481
12	54235-175810-2-SM.doc		37581	6752
13	34281-100247-5-SM.doc		27818	5359
14	46304-142650-2-SM.doc		32473	6052

شکل ۲. خروجی هر پوشه از نرم‌افزار پیکره‌ساز

پیکره با ساختاربندی مطابق استاندارد CoNLL-U [۱] است. مجموعه برچسب‌های مورد استفاده مطابق با پیکره بانک درختی PerDet موجود در سایت دادگان [۲] (رسولی، کوهستانی و مولودی ،۲۰۱۳) به‌صورت جدولی است که در ادامه آمده است.

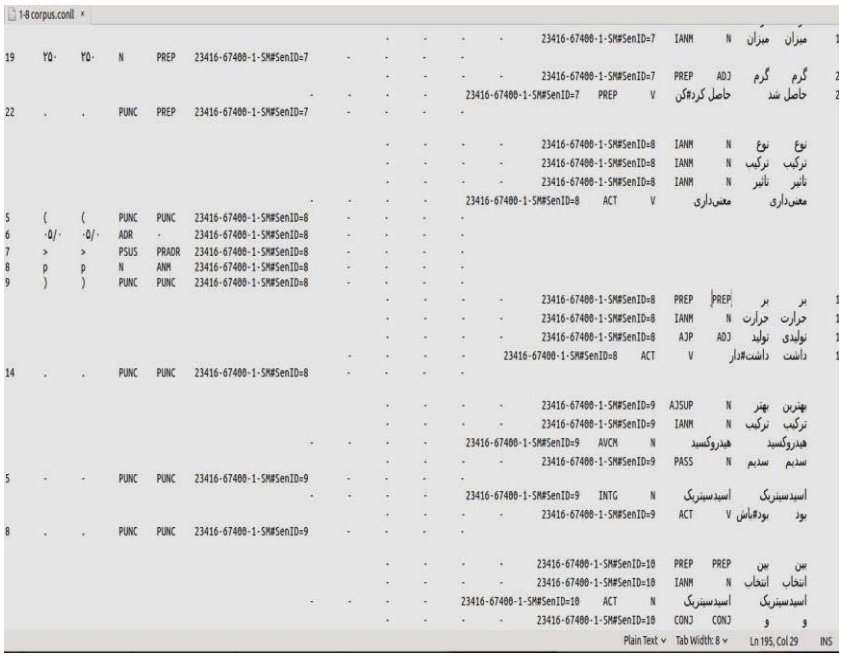
۱. توضیحات بیشتر درباره این فرمت پیکره را می‌توانید از "http://universaldependencies.org/docs/format.htm" مشاهده کنید.

2. Dadegan.ir

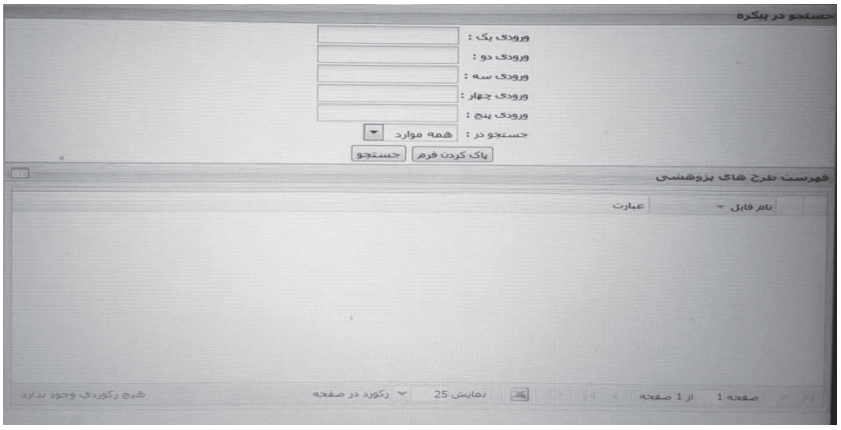
جدول ۱. برچسب‌های مورد استفاده در پیکره

Morphosyntactic features in the Persian dependency treebank				
CPOS	**FPOS**	**Person**	**Number**	**TMA**
ADJ (adjective)	AJP (positive) AJCM (comparitive) AJSUP (superlative)			
ADR (address term)	PRADR (pre-noun) POSADR (post-noun)			
ADV (adverb)	SADV (genuine)			
CONJ (coordinating conjunction)	CONJ (conjunction)			
IDEN (title)	IDEN (title)			
N (noun)	ANM (animate) IANM (inanimate)		SING (singular) PLUR (plural)	
PART (particle)	PART (particle)			
POSNUM (post-noun modifier)	POSNUM (post-noun modifier)			
POSTP (postposition)	POSTP (postposition)			
PR (pronoun)	SEPER (separate personal) JOPER (enclitic personal) DEMON (demonstrative) INTG (interogative) CREFX (common reflexive) UCREFX (noncommon reflexive) RECPR (reciprocal)	1 2 3	SING (singular) PLUR (plural)	
PREM (pre-modifier)	EXAJ (exclamatory) QUAJ (interrogative) DEMAJ (demonstrative) AMBAJ (ambiguous)			
PRENUM (pre-noun numeral)	PRENUM (pre-noun numeral)			
PREP (preposition)	PREP (preposition)			
PSUS (pseudo-sentence)	PSUS (pseudo-sentence)			
PUNC (punctuation)	PUNC (punctuation)			
V (verb)	ACT (active) PAS (passive) MOD (modal)	1 2 3	SING (singular) PLUR (plural)	See Table 3
SUBR (subordinating clause)	SUBR (subordinating clause)			

براین اساس، خروجی داده‌های پیکرهٔ حاضر به‌صورت شکل۳ است.



شکل۳. نمایی از تحلیل هر فایل



شکل۴. استخراج اطلاعات از پیکره با نرم‌افزار جانبی

۴-۶ ارزیابی دقت نرم‌افزار پیکره‌ساز

برای ارزیابی درستی ابزارهای ریشه‌یاب و برچسب‌زنی ادات سخن (نقش کلمات در

جمله) سعی شد تا حدممکن از پیکره‌های استاندارد موجود مانند پیکره PerDT (در سایت دادگان) که شامل تعداد قابل‌توجهی جملهٔ برچسب‌خورده با اطلاعات نحوی و ساخت‌واژی است (Rasooli, M. Kouhestani, and A. Moloodi، ۲۰۱۳) استفاده شود. با توجه به اینکه ریشه‌یاب ایجادشده، قابلیت ریشه‌یابی چندسطحی واژگان تا رسیدن به بن اصلی واژه را دارد، برای کاربرد موردنظر این پژوهش از ریشه‌های به‌دست‌آمده، از اولین سطح ابزار ریشه‌یابی استفاده می‌شود. همچنین در این پیکره برای برچسب نقش ادات سخن (نقش کلمات) دو سطح (با جزئیات متفاوت) وجود دارد که با توجه به کاربرد مورد نظر ابزار، سطح اول (نقش کلی کلمه در جمله) در نظر گرفته شده است.

برای محاسبهٔ درستی ابزارها، توالی واژگان ورودی به‌صورت
$W=\{w1, w2, w3, \dots, wn\}$
توالی واژگان هدف یا برچسب صحیح آنها به‌صورت
$W^*=\{w_1^*, w_2^*, w_3^*, \dots, w_n^*\}$
و توالی واژگان (یا برچسب) خروجی ابزارها به‌صورت
$W_{out}= \{w_1^{out}, w_2^{out}, w_3^{out}, \dots, w_n^{out}\}$
نامگذاری شد. برای محاسبهٔ میزان درستی ابزارها از رابطهٔ زیر استفاده می‌شود:

$$\frac{|W^* \cap W^{out}|}{|W^*|}$$

به‌طوری‌که:

$$\bigcup \{w_i^{out}\}$$

میزان درستی ابزارهای ریشه‌یابی و برچسب‌زنی نقش لغات تفکیک عملکرد برای افعال (کلمات دارای نقش فعل) و کلمات غیرفعل و به‌صورت کلی، در جدول زیر ذکر شده است:

جدول۲. میزان درستی ابزارهای پردازش زبان طبیعی (متن) فارسی

نام ابزار	پیکره تست	میزان صحت (افعال)	میزان صحت (غیرفعل)	میزان صحت کلی
ریشه‌یاب (سطح اول)	دادگان	94.36٪	92.66٪	93.24٪
برچسب‌زنی نقش کلمات (سطح اول)	دادگان	97.8٪	96.09٪	96.8٪

خلاصه و نتیجه‌گیری

امروزه نیاز به افزودن فراداده به داده‌های عظیم جهت پردازش آنها به‌وسیلهٔ ماشین، بیش از پیش احساس می‌شود. یکی از بهترین روش‌ها برای این نوع کاربرد، استفاده از پیکره است. برای تهیهٔ پیکره، متون مورد استفاده باید به‌صورتی دربیاید که به وسیلهٔ ماشین قابل خواندن باشد. در این پژوهش ابزار پیکره‌ساز طراحی شد و سپس از ۱۱۰۰ مقاله علمی‌پژوهشی پیکره متون علمی ساخته شد. از ساده‌ترین موارد کاربرد این پیکره، بررسی داده‌ها با شمارش جمله‌ها، واژه‌ها و تکواژهاست. استخراج کلیدواژه‌ها و واژه‌های همایند نیز از این قبیل است. به‌طور دقیق، حوزه‌های مورد هدف این پیکره مطالعات زبان برای اهداف خاص و دانشگاهی است. به‌عنوان مثال، بر روی واژگان رشته‌های دانشگاهی، همنشینی واژگان، زنجیره‌های واژگانی و روابط دستوری است. همچنین با استفاده از این پیکره می‌توان فهرست‌های مختلفی از واژگان عمومی یا تخصصی گروه‌های آموزشی مختلف را استخراج کرد. این فرایند تأییدکنندهٔ آثاری چون «کاکسهد»[1] (۲۰۰۰)، «گاردنر و دیویس»[2] (۲۰۱۳) است. همچنین با استفاده از این پیکره می‌توان شیوهٔ نگارش متون علمی در بین رشته‌ها و نویسندگان مختلف را بررسی کرد؛ از جمله آثاری که تأییدکنندهٔ این مطلب «هایلند و تسو»[3] (۲۰۰۴) است. از دیگر کاربردهای

1. Coxhead
2. Gardner & Davies
3. Hyland & Tse

این پیکره، استفاده برای ساخت هستان‌شناسی‌های مختلف است؛ درست همان‌گونه که «مرادی، وزیرنژاد و بحرانی» (۱۳۹۴) از سه پیکره همشهری، بیجن‌خان و ویکی‌پدیا استفاده کرده و هستان‌شناسی دانش عرفی زبان فارسی را به‌وجود آورده‌اند.

برای بررسی موردی، میزان و نوع نشانه‌های تردید در دو گروه مورد پژوهش استخراج و مقایسه شد. نشانه‌های تردید کلماتی هستند که نشان‌دهندهٔ عدم قطعیت نویسنده از بیان گفته‌اش است و استفاده از آنها نشان‌دهنده پایبندنبودن کاملِ نویسنده به درستی و ارزش صدق یک گزاره است. از آنجا که دیدگاه نویسنده در تمام جملات وجود دارد، لازم است ادعاهای نویسندگان مقاله‌های علمی بسیار دقیق، با احتیاط و همراه با تواضع بیان شود تا مقبول طبع مخاطبان قرار گیرد. نشانه‌های تردید، واسط اطلاعات متن و تفسیرنویسنده است. هدف از بررسی نشانه‌های تردید، بررسی تأثیر حوزه پژوهش بر روی راهکارهای مورداستفاده در نشانه‌های تردید توسط اعضای هیئت علمی دانشگاه فردوسی مشهد در نگارش مقاله‌های علمی و همچنین آزمایش پیکره، ایجاد شده است. قبل از این، پژوهشگران دیگری به بررسی عبارت‌های احتیاط‌آمیز پرداخته بودند، اما کارهای انجام‌گرفته به‌صورت دستی و بر روی تعداد محدودی مقاله یا پایان‌نامه انجام گرفته بود. همچنین از دسته‌بندی‌های دیگری برای تحلیل خود استفاده کرده بودند. نتایج حاصل از این پژوهش به‌علت انبوه‌بودن داده‌های مورد استفاده، قابلیت تعمیم به‌عنوان الگو در نوشتن مقاله‌های علمی را دارا هستند که نتایج دقیق آن در اثر دیگری در حال انتشار است. نتایج این تحقیق ثابت می‌کند پیکرهٔ ساخته‌شده در این پژوهش منبع بسیار خوبی برای انجام سایر تحقیقات بر روی متون علمی است.

به‌طور خلاصه می‌توان گفت این پژوهش در پنج مرحله انجام شده است. نخست، تبدیل متون به پیکره خام اولیه به فرمت قابل خواندن برای ماشین. دوم، گردآوری منابع دستور زبان برای کار بر روی پیکره آغازین. در مراحل سوم و چهارم انواع برچسب به پیکره اضافه شد و در مرحله پنجم استخراج دانش صریح از روی پیکره امکان‌پذیر گردید. اکنون پیکره آمادهٔ استخراجِ اطلاعاتِ ضمنی در سطوح مختلف از جمله بررسی ریشهٔ

کلمات یا مقولهٔ کلمات توسط کاربران است. امکان استفادهٔ محدود از نرم‌افزار از طریق ارسال درخواست به آدرس ایمیل نویسنده نیز مسئول وجود دارد. همچنین دسترسی آزاد و نامحدود به پیکره در آیندهٔ نزدیک فراهم می‌شود.

منابع

- اُگرادی، ویلیام؛ دابروولسکی، مایکل و آرنف، مارك (۱۳۸۰). *درآمدی بر زبان‌شناسی معاصر*، ترجمهٔ علی درزی، تهران: سمت.
- دانشکار آراسته، پویان (۱۳۸۱). نرم‌افزار تشخیص فعل در زبان فارسی. پایان‌نامه کارشناسی ارشد، تهران: دانشگاه علامه طباطبایی.
- عاصی، مصطفی. (۱۳۸۳). «پردازش دستوری زبان فارسی با استفاده از رایانه»، *نامه فرهنگستان*، ۱ (۱)، صص ۵۱- ۲۹.
- فرخ، ماندانا (۱۳۸۱). بررسی ساختمان افعال ساده و مرکب فارسی و تدوین روش‌های سرواژه‌سازی به کمک رایانه، پایان‌نامه کارشناسی ارشد، تهران: دانشگاه تهران.
- مرادی، مهدی؛ وزیرنژاد، بهرام و بحرانی، محمد (۱۳۹۴). «ساخت هستان‌شناسی دانش عرفی زبان فارسی با رویکردی تلفیقی»، *پژوهشنامه پردازش و مدیریت اطلاعات*، (۱)۳۱، صص۱۲۴- ۱۰۹.
- مشهور، پروین‌دخت و فقیری، غلام‌محمد (۱۳۹۱). «بررسی و تحلیل انواع زمان در دفتر اول، دوم و سوم مثنوی مولوی با رویکرد سبک شناسی رایانشی .پیکره‌ای»، *مجله زبانشناسی و گویش‌های خراسان*، دانشگاه فردوسی مشهد، ۶، صص ۹۹-۷۹.
- یاری، علیرضا (۱۳۹۴). بررسی پیکره‌ها و ابزارهای پردازش زبان طبیعی. (گزارش طرح) پژوهشکده ارتباطات و فناوری اطلاعات.

- Abdul-Mageed, M., Diab, M. and Kübler, S. (2014). SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, **28**(1): p. 20-37.
- Baker, P. (2006). *Using corpora in discourse analysis*. Continuum Discourse Series.
- Bijankhan, M., (2004). The role of the corpus in writing a grammar: An introduction to a software. *Iranian Journal of Linguistics*, 19(2).
- Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*(2), 213-238.
- A. Kiryakov, B. Popov, D. Manov and D. Ognyanoff, (2004). Semantic Annotation, Indexing, and Retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, pp. 49- 79.
- Feely, W., Manshadi, M. Frederking, R. and Levin, L. (2014). The CMU METAL Farsi NLP Approach, in *Proceedings of the Ninth International*

Conference on Language Resources and Evaluation (LREC'14), pp. 4052-4055.

- The CMU METAL Farsi NLP Approach. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14).
- Ganchev, K. Taskar, B. Pereira, F. and Gama, J. (2009). Posterior vs parameter sparsity in latent variable models, *Advances in Neural Information Processing Systems*.
- Gardner, D., & Davies, M. (2013). A New Academic Vocabulary List. *Applied Linguistics*.
- Granger, S., Gilquin, G., & Meunier, F. (2015). *The Cambridge Handbook of Learner Corpus Research*: Cambridge University Press.
- Hyland, K., & Tse, P.M. 2004. Meta discourse in academic writing: A reappraisal. *Applied Linguistics*, *25*(2), 156-177.
- Khallash, M. and M. Imany. (2014). Hazm: Python library for digesting Persian text. [cited 2015; Available from: https://github.com/sobhe/hazm.
- Jadidinejad, A.H., F. Mahmoudi, and J. Dehdari, (2010). Evaluation of PerStem: a simple and efficient stemming algorithm for Persian, in Workshop of the Cross-Language Evaluation Forum for European Languages. *Springer*. p. 98-101.
- Manshadi. M. (2013). *Farsi Verb Tokenizer*. Available: https://github.com/mehdi-manshadi/Farsi-Verb-Tokenizer.
- Manshadi, M. (2015). Farsi Verb Tokenizer. 2013. [cited 2015; Available from: https://github.com/mehdi-manshadi/Farsi-Verb-Tokenizer.
- Megyesi, B. (1999). Improving Brill's POS tagger for an agglutinative language. in Proceedings of the Joint SIGDAT Conference on Empirical Methods, *Natural Language Processing and Very Large Corpora*.
- McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An Introduction*: Edinburgh University Press.
- Michelson, Matthew, and Craig A. Knoblock. (2005). Semantic annotation of unstructured and ungrammatical text, *International Joint Conference on Artificial Intelligence*. Vol. 19. Lawrence Erlbaum Associates LTD.
- Talantikitr, H. N., Aissani, D. and Boudjlida, N. (2005). Semantic annotations for web services discovery and composition. *Computer Standards & Interfaces,* 31(6), pp. 1108- 1170.
- Rasooli, M.S., et al. (2011). A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank, *5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*.
- Rasooli, M.S., M. Kouhestani, and A. Moloodi. (2013). Development of a Persian syntactic dependency treebank. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*
- Sarabi, Z., H. Mahyar, and M. Farhoodi. (2013). ParsiPardaz: Persian Language Processing Toolkit. *Computer and Knowledge Engineering (ICCKE), 3th International eConference on*. IEEE.
- Seraji, M., B. Megyesi, and J. Nivre. (2012). A basic language resource kit for

Persian. *Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey*. European Language Resources Association.

- Seraji, M. (2013). PrePer: A Pre-processor for Persian, *Proceedings of The Fifth International Conference on Iranian Linguistics (ICIL5)*, Bamberg, Germany.
- Shamsfard, M., (2011). Challenges and open problems, *Persian text processing.* Proceedings of LTC.
- Shamsfard, M., H.S. Jafari, and M. Ilbeygi. (2010). STeP-1: A Set of Fundamental Tools for Persian Text Processing. *LREC*.
- Teubert, W. (1999). Corpus Linguistics: A Partisan View, *International Journal of Corpus Linguistics*, 4(1), 1-10.
- Usbeck, Ricardo, et al. (2015). GERBIL: general entity annotator benchmarking framework. *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.
- Lin, Yun. (2008). Semantic annotation for process models: Facilitating process knowledge management via semantic interoperability.