علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۵، شماره ۱، ۱۳۹۶
صفحات ۲۲- ۳۰
مقاله عادی

نشریه علوم رایانش و فناوری اطلاعات

دگربیان: توسعه پیکره متنی فارسی جملات و عبارات معادل به کمک روش جمع‌سپاری

رضا معانی‌جو سید ابوالقاسم میرروشندل

دانشکده فنی، دانشگاه گیلان، رشت، ایران

چکیده

جملات و عبارات دِگَربَیان، بیانی متفاوت از مفهومی یکسان هستند. شناسایی دگربیان‌ها یکی از وظایف مهم سامانه‌های پردازش زبان طبیعی است. با وجود اهمیت این موضوع، پیکره عبارات دگربیان در زبان فارسی توسعه‌نیافته است. هدف این مقاله ارائه روشی جهت تهیه پیکره عبارات و جملات دگربیان در زبان فارسی است. به همین منظور سامانه‌ای خودکار و بی‌ناظر جهت استخراج عبارات و جملات دگربیان ارائه می‌شود که از داده‌های دریافت شده از خبرهای خبرگزاری‌ها استفاده می‌کند. با استفاده از الگوریتمی مبتنی بر معیار جاکارد، نمونه‌های دگربیان در دو سطح عبارات و جملاتی با اندازه‌های متفاوت استخراج می‌شوند. سپس نمونه‌های به دست آمده به کمک تکنیک‌های جمع‌سپاری و سامانه‌ای که تحت پیام‌رسان تلگرام پیاده‌سازی شده، نشانه‌گذاری شده و در سه رده دگربیان، تقریباً دگربیان و نامرتبط ارائه می‌شوند و نتایج حاصل‌شده ارزیابی می‌گردند. در حال حاضر تعداد ۱,۵۲۳ نمونه نشانه‌گذاری شده در نسخه ۱,۰ از پیکره موجود است که در دسترس عموم است.

کلمات کلیدی: پردازش زبان‌های طبیعی، پیکره، جمع‌سپاری، روش‌های بی‌ناظر، عبارات دگربیان، معیار فاصله.

۱- مقدمه

جملات و عباراتی که مفهوم یکسانی را به شکل متفاوتی ارائه می‌نمایند، دگربیان[۱] خوانده می‌شوند. شناسایی این جملات یکی از وظایف بسیار مهم در پردازش زبان‌های طبیعی است که در بسیاری از سامانه‌های بازیابی اطلاعات[۲]، خلاصه‌سازی متن[۳] و شناسایی سرقت ادبی[۴] مورداستفاده قرار می‌گیرد. در این کاربردها نیاز است تا با جملات دگربیان به نحو یکسانی رفتار شود. این جملات دارای انواع متفاوتی هستند که باعث می‌شود تا شناسایی آن‌ها به آسانی صورت نگیرد [۱، ۲].

اغلب روش‌هایی که به منظور شناسایی جملات دگربیان مورداستفاده قرار می‌گیرند، نیازمند استفاده از داده‌هایی به شکل جفت جملات مشابه هستند [۳] که باید به صورت پیکره‌ای مناسب در دسترس باشند. در زبان انگلیسی پیکره‌های بسیاری به این منظور توسعه یافته است که از آن میان می‌توان به پیکره جملات دگربیان شرکت مایکروسافت اشاره کرد [۳]. این پیکره یکی از معروف‌ترین و پرکاربردترین پیکره‌های موجود است که به صورت خودکار و از میان خبرهای به دست آمده در سطح اینترنت ساخته شده است. جملات، به کمک روش فاصله ویرایشی[۵] و با استفاده از دو جمله ابتدایی هر خبر استخراج شدند و در اختیار داوران انسانی قرار گرفتند تا نشانه‌گذاری شوند.

ایده اصلی این است که جملات ابتدایی هر خبر، چکیده‌ای از مطالب آن را به‌طور خلاصه بیان می‌کند؛ بنابراین تنها با بررسی این جملات از میان خبرگزاری‌های مختلف، می‌توان جملات مشکوک به دگربیان را استخراج نمود. به این ترتیب نمونه‌های موجود در دو دسته دگربیان و غیردگربیان تقسیم‌بندی می‌شوند که در مجموع شامل ۵,۸۰۱ جفت جمله است که در دو بخش نمونه‌های آموزشی و آزمایشی ارائه شده‌اند و تاکنون روش‌های مختلفی جهت شناسایی دگربیان‌ها بر روی آن اعمال شده است.

داده‌های حجیم شبکه‌های اجتماعی، منبع مناسب دیگری برای استخراج عبارات دگربیان هستند. توییتر[۶] یکی از این منابع است که به صورت گسترده برای استخراج دگربیان‌ها مورداستفاده قرار گرفته است [۴، ۵]. کاربران توییتر در بسیاری از موارد در مورد موضوعات یکسانی گفتگو می‌کنند که می‌توانند برای استخراج دگربیان‌ها به کار گرفته شود. در این روش‌ها، تکنیک‌های مختلفی مانند بررسی ویژگی‌های همپوشانی کلمات و حروف مورداستفاده قرار می‌گیرد تا دگربیان‌ها استخراج گردند. موضوع توییت‌ها و جداسازی نوشته‌های هم‌موضوع می‌تواند روند تشخیص عبارات دگربیان را به خوبی پیش برد. هشتک‌های مطالب در توییتر به منظور شناسایی موضوع توییت‌ها به کار می‌رود. در این راستا با بررسی همپوشانی کلمات نوشته‌هایی که موضوع یکسانی (با توجه به هشتک) دارند، می‌توان جملات دگربیان را شناسایی کرد.

در روش دیگری که بر روی زبان روسی انجام شده است، به‌طور مشابه از خبرهای خبرگزاری‌ها استفاده شده است و از روش جمع‌سپاری[۷] به منظور نشانه‌گذاری نمونه‌های به دست آمده استفاده می‌شود. همچنین نمونه‌ها در سه دسته نامرتبط، دگربیان و مرتبط قرار گرفته‌اند [۶]. برای شناسایی نمونه‌ها در این روش، از یک معیار شباهت‌سنجی معنایی مرتبط با ماتریکس شباهت استفاده شده است. پیکره‌های دیگری نیز به کمک بازنویسی غیرخودکار عبارات موجود توسط افراد [۷] و یا استخراج خودکار از میان پیکره‌های موازی دو زبانه تولید شده‌اند [۸] که همگی در زبان‌های غیرفارسی هستند.

با وجود اهمیت پیکره‌های عبارات دگربیان، تاکنون پیکره‌ای فارسی که در دسترس عموم باشد، توسعه نیافته است. به همین منظور نیاز به توسعه پیکره‌ای جهت شناسایی جملات دگربیان در زبان فارسی احساس می‌شود تا بتواند در کاربردهای مختلف استفاده شود. با این حال پیکره‌های مشابه و خاص منظوره به خصوص در رابطه با تشخیص سرقت ادبی توسعه یافته‌اند که از آن میان می‌توان به پیکره فارسی ارزیابی سامانه‌های تقلب‌یاب اشاره کرد [۹]. این پیکره بر مبنای کپی‌برداری‌های موجود در صفحات فارسی ویکی‌پدیا ساخته شده و حاوی ۱,۵۰۰ نوشته است که از آن میان ۴۱۱ مورد دارای شکل‌هایی از تقلب به صورت جابجایی، حذف، اضافه شدن و جایگزینی کلمات هم‌معنا است که با درجه‌بندی‌های مختلف از نظر میزان سرقت ادبی همراه شده است و محدود به سطح جملات و عبارات نبوده و نمی‌تواند به عنوان پیکره‌ای جهت شناسایی جملات دگربیان به کار رود.

علاوه بر این پیکره‌هایی با رویکرد ترجمه خودکار نیز توسعه یافته‌اند و هدف اصلی آن‌ها توسعه و ارزیابی سامانه‌های ترجمه خودکار است که از آن میان می‌توان به پیکره موازی انگلیسی- فارسی میزان و پیکره موازی انگلیسی- فارسی پیام اشاره کرد [۱۰، ۱۱]. کاربرد این پیکره‌ها نیز در جهت ترجمه ماشینی است و اغلب به صورت چندزبانه و موازی هستند.

هدف این پژوهش، پر کردن خلأ پیکره‌ای تخصصی در زبان فارسی برای جملات و عبارات دگربیان است. نسخه ۱,۰ از این پیکره در حال حاضر شامل ۱,۵۲۳ نمونه از جملات و عبارات بوده که درجه‌های مختلفی از یکسانی معنایی را دارا هستند و پس از شناسایی خودکار، توسط یک روش جمع‌سپاری نوآورانه و کمک گرفتن از داورهای انسانی، نشانه‌گذاری شده است. عبارات منتخب شامل قسمت‌هایی از خبرهای دریافت شده خبرگزاری‌های مختلف بوده و طیف وسیعی از موضوعات خبری را تحت پوشش قرار می‌دهد.

در ادامه مقاله پیشرو، ابتدا تعریفی از عبارات و جملات دگربیان را ارائه خواهیم کرد که به درک بهتر مطالب و موضوع مقاله، کمک می‌کند. در بخش سوم، روش پیشنهادی به منظور ایجاد خودکار پیکره جملات و عبارات دگربیان، به همراه تشریح داده‌های خام ارائه می‌شود و پس از توصیف روش‌های نشانه‌گذاری و ارزیابی پیکره ایجادشده، نتیجه‌گیری‌ها و کارهای آتی بیان می‌شوند.

۲- عبارات و جملات دگربیان

تعریف دقیق عبارات دگربیان کار دشوار و بحث‌انگیزی است و با توجه به کاربرد موردنیاز متفاوت خواهد بود. در اغلب تعاریف ارائه شده توسط زبان‌شناسان، یکسانی کامل معنی دو عبارت را در نظر نمی‌گیرند و اجازه می‌دهند تا بتوان عباراتی که تقریباً یکسان هستند را نیز دگربیان بنامیم. تقریبی بودن یکسانی معانی باعث می‌شود که در شناسایی آن‌ها دچار ابهام شویم. تعاریف زیادی برای دگربیان ذکر شده است امّا هیچ کدام نتوانسته‌اند تا ابهامات موجود برای شناسایی آن‌ها را به‌طور کامل برطرف سازند. در واقع همین ابهامات هستند که باعث دشواری در پیاده‌سازی سیستم‌های شناسایی دگربیان شده‌اند.

اگر چه تعریف دقیقی در مورد عبارات دگربیان وجود ندارد که خالی از ابهام باشد، امّا دسته‌بندی‌های مختلفی برای الگوهایی که دگربیان‌ها در آن‌ها یافت می‌شوند، ارائه شده است [۱۲]. به عنوان مثالی از این الگوها می‌توان جایگذاری کلمات مترادف در عبارت، کوتاه‌سازی جملات و تغییر ساختار جمله را ذکر نمود که توسط سیستم‌های شناسایی و تولید دگربیان‌ها مورداستفاده قرار می‌گیرد.

در بسیاری از موارد نیازمند یکسانی دقیق دو عبارت یا جمله از نظر معنایی نیستیم و نزدیک بودن مفهوم دو عبارت کافی است. از طرف دیگر کاربردهای دیگری نیز وجود دارند که باید شکل یکسانی را در رابطه با مفاهیم به کار گرفته شده، در نظر داشته باشند و حتی وجود حقایق بیشتر در یکی از عبارات باعث می‌شود تا دو عبارت را از نظر معنایی، معادل ندانیم [۲]. به عنوان مثال در هنگام خلاصه‌سازی متن، یکسانی دقیق مدنظر نیست در حالی که در رابطه با سامانه‌های پرسش و پاسخ این یکسانی در نظر گرفته می‌شود [۶]. به عنوان مثال، نمونه‌های زیر را در نظر بگیرید:

- گفت‌وگوی بیش از ۱۹۰ کشور جهان برای تنظیم یک پیمان جدید آب و هوایی جهت جلوگیری از گرمایش زمین، روز شنبه با توافق نمایندگان این کشورها در پاریس به سرانجام رسید.
- نخستین توافق بین‌المللی که با هدف جلوگیری از افزایش دمای کره زمین به بیش از دو درجه سلسیوس، به امضای نمایندگان نزدیک به ۲۰۰ کشور دنیا رسید، از روز گذشته به مرحله اجرا درآمد.

این دو جمله در مورد موضوع مشخص و یکسانی بحث می‌کنند و در اکثر نیازها می‌توان آن‌ها را معادل هم دانست؛ اما با دقت در منطق و موارد ذکر شده می‌توان متوجه شد که در جمله دوم به اجرایی شدن توافق نیز اشاره شده که در جمله اول ذکر نشده است. اینکه جملات این چنینی را دگربیان در نظر بگیریم یا خیر، موضوعی است که با توجه به کاربرد متفاوت خواهد بود. گاهی نیز چنین نمونه‌هایی را تقریباً دگربیان می‌دانند [۶].

در این مقاله، برخلاف بسیاری از روش‌های پیشین راه‌حلی بینابینی در نظر گرفته می‌شود تا کاربردهای متفاوت در نظر گرفته شوند. برای رسیدن به این هدف، فرایند نشانه‌گذاری نمونه‌ها با در نظر گرفتن یکسانی کامل معنای دو عبارت، در کنار مرتبط بودن یا نبودن آن‌ها انجام می‌شود و سه رده مختلف جهت نشانه‌گذاری نمونه‌ها در نظر گرفته می‌شود.

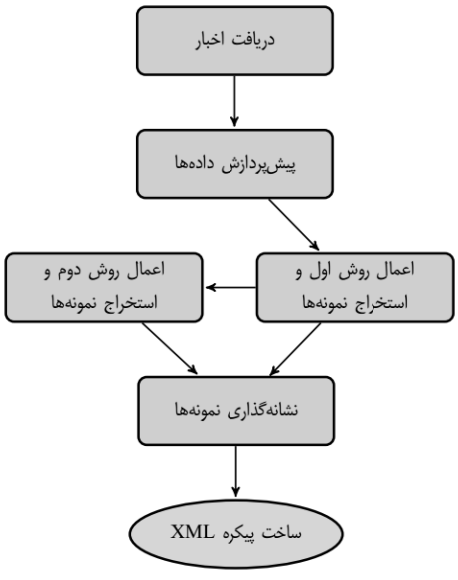
مزیت این راهکار در این است که پیکره‌ای غنی و متشکل از انواع حالات ایجاد می‌گردد که مناسب طیف بیشتری از کاربردها خواهد بود. در صورتی که تنها به دو رده نیاز باشد، می‌توان رده دگربیان و تقریباً دگربیان را در یک رده گنجاند و پیکره توسعه‌یافته همچنان قابل استفاده خواهد بود.

۳- روش پیشنهادی

در این قسمت روش‌های مختلف به منظور ایجاد خودکار پیکره جملات و عبارات دگربیان، به همراه مراحل و جزئیات آن‌ها ارائه می‌شود. الگوریتم‌های شناسایی عبارات دگربیان با روش‌های ساخت پیکره عبارات دگربیان متفاوت هستند. در روش‌هایی که جهت ساخت پیکره به کار می‌روند، یافتن نمونه‌هایی از تمامی رده‌ها در نظر گرفته می‌شود تا پیکره نهایی شامل تمامی رده‌ها باشد در حالی که شناسایی عبارات دگربیان بر روی تنها یک رده از آن‌ها تمرکز می‌کند. روش پیشنهادی، تلفیقی از روش‌های انجام شده در پیکره مایکروسافت و روش‌های مشابه با آن بوده و هدف آن ساخت پیکره به همراه سه رده مختلف است که عبارات کاملاً هم‌معنی (دگربیان) و تقریباً هم‌معنی را جداگانه در نظر گیرد [۳، ۶]. یکی از ویژگی‌های پیکره ساخته‌شده این است که محدود به سطح جمله نبوده و شامل عبارات دگربیان نیز هست که به ویژه در کاربردهایی که با عبارات پرس‌وجو سروکار دارند، بسیار مفید خواهد بود.

علاوه بر این با توجه به ویژگی‌هایی که زبان فارسی و داده‌های خبری جمع‌آوری شده دارا هستند و اطلاعات اضافی مانند عناوین خبری که در دسترس قرار دارد، می‌توان روند تشخیص این عبارات را بهبود داد که در ادامه بحث، توضیح داده می‌شود.

شکل ۱ چارچوب کلی روش پیشنهادی را نشان می‌دهد. در ابتدا لازم است که خبرهای خبرگزاری‌ها به کمک خزنده‌های[۸] وب استخراج گردند. این خبرها پیش‌پردازش می‌شوند و سپس جملاتی که احتمال می‌رود تا دگربیان باشند، توسط دو روش مختلف استخراج می‌گردند که باید به ترتیب اجرا شوند. جملات به دست آمده توسط این دو روش، به کمک روش جمع‌سپاری، نشانه‌گذاری می‌شوند و به این ترتیب پیکره مورد نظر تولید می‌گردد. در ادامه این بخش، هر یک از مراحل تولید پیکره عبارات دگربیان را مورد بررسی قرار می‌دهیم.



شکل ۱- چارچوب کلی روش پیشنهادی

۳-۱- داده‌های مورداستفاده

روش‌های استخراج عبارات دگربیان را می‌توان به دو حالت غیرخودکار و خودکار تقسیم نمود. در روش‌های غیرخودکار، نمونه‌های دگربیان به‌طور مستقیم توسط عامل انسانی وارد پیکره می‌شوند و الگوریتم‌ها و برنامه‌های کامپیوتری نقش کمتری را بر عهده دارند. به عنوان مثال با شروع از یک جمله یا عبارت، از افراد خواسته می‌شود که آن‌ها را به صورت متوالی بازنویسی کنند و هریک از این بازنویسی‌ها به عنوان نمونه‌هایی در داخل پیکره قرار می‌گیرند. به پیکره‌هایی که به این شکل و به‌طور مستقیم توسط عامل انسانی تولید می‌شوند، مصنوعی نیز گفته می‌شود. با وجود اینکه پیکره‌ای که به این روش به دست می‌آید دارای کیفیت بالایی است، روش‌های غیرخودکار بسیار زمان‌گیر و پرهزینه هستند و در بسیاری از موارد روش‌های خودکار به آن‌ها ترجیح داده می‌شوند [۷، ۱۳].

در صورتی که روش ساخت پیکره به صورت خودکار باشد، نوع منبع داده اهمیت پیدا می‌کند. منابعی که در ساخت چنین پیکره‌هایی به کار می‌روند باید از نظر ذاتی امکان وجود عبارات معادل را فراهم نمایند به نحوی که احتمال یافتن نمونه‌ها افزایش یابد. در غیر این صورت یافتن عبارات دگربیان دشوار خواهد بود تا حدی که ممکن است در یک حجم داده‌ای بزرگ، تنها چند نمونه یافت شود. به عنوان مثال، ترجمه‌های یکسانی که از نوشته‌های مختلف به دست می‌آید و یا وجود نوشته‌هایی که در رابطه با موضوع یکسانی مانند یک خبر به خصوص هستند، احتمال یافتن عبارات دگربیان را افزایش می‌دهند. مزیت روش‌های خودکار، سرعت بالا و هزینه کم آن‌ها است که با کیفیت مطلوب نمونه‌های بدست آمده، همراه شده است [۳، ۱۴، ۱۵].

مجموعه داده‌ای که در این پژوهش مورداستفاده قرار می‌گیرد شامل خبرهای به دست آمده در بازه زمانی ۲ ماه از مجموع بیش از ۱۱ خبرگزاری مختلف است که به کمک یک ابزار خزنده، دریافت شده است. ساختار صفحات در هر یک از این خبرگزاری‌ها با دیگری متفاوت است. صفحات می‌توانند شامل فهرست‌ها، تبلیغات و بخش‌های مختلفی باشند که تنها خود خبر به همراه اطلاعات مربوط به آن مورد نیاز است؛ بنابراین برای دریافت اطلاعات از هر یک از این خبرگزاری‌ها لازم است که ابزار خزنده به صورت مجزا بر روی هر از این خبرگزاری‌ها توسعه یابد و اطلاعات مورد نیاز را دریافت نماید.

نوشته‌های خبری‌ای که در رابطه با موضوعات یکسان نوشته می‌شوند به صورت ذاتی می‌توانند شامل عبارات و جملات معادل هم باشند. اطلاعات مختلف هر خبر، مانند عنوان، متن خبر، برچسب‌های هر خبر، موضوع و موارد دیگر برحسب مورد از میان صفحات HTML به دست آمده و ذخیره می‌شوند تا بعداً به منظور شناسایی دگربیان‌ها مورداستفاده قرار گیرند. بسیاری از این اطلاعات ناقص بوده و یا در تمامی خبرگزاری‌ها موجود نیستند. به عنوان مثال ویژگی برچسب موضوعی هر خبر توسط همه خبرگزاری‌ها و یا برای تمامی خبرها پیاده‌سازی نشده است و برای بسیاری از خبرها موجود نیست پس باید در انتخاب ویژگی‌هایی که در الگوریتم به کار می‌روند دقت شود.

۳-۲- پیش‌پردازش داده‌ها

از اولین کارهایی که در تهیه پیکره انجام می‌شود مرحله پیش‌پردازش و یکسان‌سازی متن است. یکی از ویژگی‌های خوب داده‌های مورداستفاده این است که خبرهای تهیه شده توسط خبرگزاری‌ها، قبل از انتشار مورد بازبینی قرار می‌گیرند و دارای حداقل ایرادات نگارشی و املایی بوده و از نظر نوع بیان هم به صورت رسمی هستند. با این حال، همچنان امکان دارد که خطاهایی در آن‌ها وجود داشته باشد. این خطاها می‌توانند شامل انواع خطاهای نگارشی و املایی باشند. علاوه بر این ممکن است در هنگام دریافت خبرها توسط خزنده، اطلاعات ناخواسته‌ای مانند داده‌های مربوط به تصاویر، برچسب‌های صفحات HTML و تبلیغات نیز جمع‌آوری شوند که باید حذف گردند. همچنین شکل‌های متنوعی که برای هر یک از لغات و کلمات وجود دارد، باعث می‌شود که الگوریتم‌های پردازش زبان به خوبی عمل نکنند. به عنوان مثال برای کلمه‌ای مانند کتاب، اشکال

مختلفی مانند کتابی و کتاب‌ها نیز وجود دارد که به مفهوم یکسانی اشاره می‌کنند و باید همسان در نظر گرفته شوند. در مرحله پیش‌پردازش، تمامی ایرادات موجود در داده‌های به دست آمده برطرف می‌گردند و یکسان‌سازی‌ها انجام می‌شوند.

به منظور حذف ایرادات نگارشی و املایی از ابزارهای هضم [۱۶] و ویراست‌یار [۱۷] استفاده شده است که به صورت بازمتن ارائه شده‌اند و برای پردازش متون فارسی مورداستفاده قرار می‌گیرند. بعد از حذف اطلاعات اضافی و زائد از خبرها، با استفاده از ویراست‌یار ایرادات نگارشی و املایی موجود در خبرها برطرف می‌گردد. در مرحله بعدی با استفاده از هضم، یکسان‌سازی نویسه‌ها صورت می‌گیرد و کلمات معروف به کلمات ایست[۹] حذف می‌گردند. کلمات ایست، کلماتی هستند که به صورت مکرر در متن تکرار می‌شوند و از نظر بار معنایی چیزی به جملات اضافی نمی‌کنند و در کاربردهای شباهت‌سنجی و پردازش زبان، به‌راحتی قابل چشم‌پوشی هستند. بعد از انجام این مراحل، داده‌های پردازش شده به عنوان ورودی به الگوریتم‌های شناسایی عبارات دگربیان تحویل داده می‌شوند.

۳-۳- شناسایی خودکار به کمک معیار فاصله

به‌طور معمول، شناسایی خودکار دگربیان‌ها به کمک شباهت‌سنجی متنی دو عبارت یا جمله صورت می‌گیرد. معیارهای مختلفی جهت مقایسه و شباهت‌سنجی رشته‌های متنی وجود دارد. در حالی که روش‌های مختلف شباهت‌سنجی مانند TF-IDF[۱۰] به منظور مقایسه معنایی دو متن به کار می‌روند، روش‌های ساده‌تری نیز وجود دارند که شباهت دو متن در آن‌ها تنها از نظر رشته‌ای بررسی می‌شوند. یکی از روش‌هایی که به صورت عمده جهت استخراج عبارات دگربیان به کار رفته است، محاسبه فاصله ویرایشی دو متن را در نظر می‌گیرد و مبتنی بر این فرض است که عبارات معادل دارای فاصله ویرایشی اندکی نسبت به یکدیگر هستند [۳]. طبق تعریف، فاصله ویرایشی به حداقل تعداد حرکات جایگزینی، درج و حذف برای تبدیل یک رشته کاراکتری به رشته کاراکتری دیگر اطلاق می‌شود. این معیار به‌خصوص در زمانی که دو رشته از نظر اندازه نزدیک به‌هم بوده و بیش از حد هم کوتاه نباشند، به‌خوبی کار می‌کند و به عنوان روشی کارا در زبان انگلیسی شناخته شده است.

با فرض متغیر بودن طول دو رشته و این نکته که هدف، پیدا کردن عبارات و جملات معادل باشد، پیدا کردن یک حد آستانه برای معیار فاصله ویرایشی، کار دشواری خواهد بود علاوه بر این، جابجایی ترتیب کلمات که در زبان فارسی نیز بسیار رایج است، بر سختی کار خواهد افزود. روش پیشنهادی‌ای که در این پژوهش برای حل این مشکل ارائه شده، استفاده از معیار فاصله جاکارد[۱۱] است. در این روش دو عبارت به صورت مجموعه‌ای از کلمات درنظر گرفته می‌شوند؛ بنابراین ترتیب قرارگیری کلمات اهمیتی نخواهد داشت و جابجایی ترتیب کلمات، اثر منفی در الگوریتم ندارد. مزیت استفاده از فاصله جاکارد سرعت اجرای بالای آن نسبت به فاصله ویرایشی و سایر روش‌های شباهت‌سنجی متن است که با توجه به کارایی بالای آن بسیار قابل توجه است. اثربخشی روش ارائه شده به نحوی است که ما را از اعمال الگوریتم‌های پیچیده‌تر که لازمه آن‌ها جاگذاری معادل کلمات در نمونه‌ها و یا استفاده از ابزارهایی مانند وردنت[۱۲] است، بی‌نیاز می‌کند.

برای محاسبه فاصله جاکارد بین دو رشته A و B، بعد از جدا کردن هر عبارت به صورت مجموعه‌ای از کلمات، خواهیم داشت:

$$\text{Jaccard} = 1 - \frac{|A \cap B|}{|A \cup B|} \qquad (۱)$$

که نتیجه به دست آمده از این معیار، عددی بین صفر تا یک است و نزدیک بودن این عدد به صفر به معنی نزدیک‌تر بودن دو مجموعه (عبارت) به یکدیگر است. این معیار پایه و اساس روش ساخت خودکار پیکره دگربیان است که در قسمت بعدی شرح داده می‌شود.

نکته‌ای که در حین استفاده از این روش باید در نظر گرفت این است که دو عبارت باید قبل از اعمال الگوریتم، پیش‌پردازش شده و واژه‌های ایست حذف گردند و به‌جای هر کلمه ریشه آن استفاده شود تا تفاوت‌های جزئی در کلمات باعث به‌وجود آمدن اختلال در عملکرد الگوریتم نشود.

۴-۳- پیاده‌سازی استخراج خودکار عبارات دگربیان

در این قسمت روش ارائه شده جهت استخراج عبارات و جملات دگربیان ارائه می‌شود. پیاده‌سازی این روش بر مبنای روش استخراج بی‌ناظر و به کمک معیار فاصله جاکارد است که در قسمت قبل شرح داده شد. داده‌های خام شامل خبرهایی است که به صورت روزانه جمع‌آوری شده است و به‌همین دلیل احتمال زیادی وجود دارد که خبرهایی یکسان، توسط خبرگزاری‌های متفاوت پوشش داده شده باشند و مفاهیم یکسانی در آن‌ها تکرار شود. هدف اصلی، استخراج عبارات و جملات دگربیان از میان این خبرهای یکسان است. همچنین همان‌طور که پیش‌تر در قسمت ۲ اشاره کردیم، علاوه بر جملات، استخراج عبارات نیز مدنظر قرار می‌گیرد.

استخراج نمونه‌ها توسط دو روش مختلف پیاده‌سازی شده‌اند. ابتدا برای افزایش دقت معیار فاصله جاکارد و یافتن سریع‌تر نمونه‌های دگربیان، دامنه جستجوی الگوریتم کاهش می‌یابد. این کار با تفکیک خبرهای موجود به موضوع و بازه‌های زمانی یکسان صورت می‌گیرد. جهت تفکیک زمانی، خبرهایی که با هم مقایسه می‌شوند باید در روز یکسانی بررسی شوند. به همین جهت یک بازه زمانی ۲۴ ساعته برای مقایسه اخبار در نظر گرفته شده است.

بعد از تفکیک زمانی داده‌ها، عملیات جستجوی نمونه‌ها انجام می‌شود که شامل دو روشی است که در این مقاله برای استخراج نمونه‌های پیکره ارائه شده است. در روش اول، جستجو تنها میان عناوین خبرها صورت می‌گیرد. عنوان خبری معمولاً به صورت عبارت و یا جملات کوتاه است؛ بنابراین به کمک این روش قادر خواهیم بود تا نمونه‌هایی به صورت عبارات و جملات کوتاه استخراج نماییم که در روش‌های دیگر به‌دلیل عدم‌برخورد با چنین نمونه‌هایی، به سختی امکان‌پذیر بود. الگوریتم شماره ۱، مراحل انجام این روش را شرح می‌دهد. معیار جاکارد بر روی هر جفت عنوان از خبرهای خبرگزاری‌ها اعمال می‌شود و با مقایسه این مقدار با مقدار حد آستانه، نمونه‌های موردنظر استخراج می‌شوند. مقدار آستانه بالا در روش مقایسه عناوین به صورت تجربی مقدار ۰٫۷ در نظر گرفته شده است و نمونه‌هایی که مقدار پایین‌تری از این حد داشته باشند، به صورت خودکار استخراج می‌شوند. برای جلوگیری از استخراج عباراتی کاملاً یکسان و یا با تفاوت جزئی، مقادیری که به صفر نزدیک باشند (کمتر از ۰٫۱) در نظر گرفته نمی‌شوند. به عنوان نمونه‌ای از عبارات استخراج شده در این روش، خواهیم داشت:

- مصرف زیاد نمک به کبد نیز آسیب می‌زند.
- تأثیر مصرف نمک بر فعالیت کبد

در روش دوم، با استفاده از روش اول خبرهای موجود به صورت موضوعی تفکیک می‌شوند. جفت خبرهایی که عناوین آن‌ها در روش اول به عنوان دگربیان استخراج شده‌اند، انتخاب می‌شوند و تمامی جملات آن‌ها به کمک معیار جاکارد مقایسه می‌گردند. استفاده از نتایج روش اول به عنوان ورودی روش دوم، باعث می‌گردد تا دامنه جستجوی جفت جملات در الگوریتم باز هم کاهش پیدا کند و پیچیدگی محاسباتی نیز کاهش یابد که خود باعث می‌شود که جملات دگربیان را زودتر استخراج نماییم. الگوریتم شماره ۲ روند انجام کار را نشان می‌دهد. برخلاف روش اول، در این روش دیگر با عبارات مواجه نخواهیم شد و اغلب نمونه‌ها به

الگوریتم ۱ روش اول تولید پیکره دگربیان به کمک معیار جاکارد.

Input: The preprocessed news with lemmatized words.

Output: Possible paraphrase sentences and phrases and their *ID*.

```
for all ReferenceNews in preprocessed data do
    Get ReferenceNews time
    Compute ReferenceNews time period
    for all CompareNews within ReferenceNews time period do
        Compute Jaccard distance between CompareNews and ReferenceNews titles,
        if (0.1 ≤ Jaccard distance ≤ 0.7) then
            Add title of CompareNews and ReferenceNews pair to the dataset.
        end if
    end for
end for
```

الگوریتم ۲ روش دوم تولید پیکره دگربیان به کمک اعمال معیار جاکارد.

Input: The News pairs from algorithm 1 which their titles were a paraphrase.

Output: Possible paraphrase sentences and phrases and their *ID*.

```
for all ReferenceNews from News pairs do
    for all CompareNews from News pairs do
        for all Sentence pairs between CompareNews and ReferenceNews do
            Compute Jaccard distance between CompareNews and ReferenceNews sentences pair,
            if (0.2 ≤ Jaccard distance ≤ 0.7) then
                Add Sentence pair to the dataset.
            end if
        end for
    end for
end for
```

صورت جملاتی از متن خبر هستند. هدف از این روش دریافت نمونه‌هایی با طول بزرگ‌تر و به صورت جملات است تا پیکره نهایی از این لحاظ نیز غنی باشد. حد آستانه درنظر گرفته شده در این روش، نمونه‌هایی با مقدار فاصله جاکارد بین ۰٫۲ تا ۰٫۷ است. همچنین به منظور افزایش کارایی الگوریتم در صورتی که نسبت طول دو جمله بیشتر از مقدار ۱ به ۲ باشد، نمونه مربوط در نظر گرفته نمی‌شود. مثالی از روش دوم به صورت زیر است:

- «کمال خرازی» وزیر خارجه اسبق جمهوری اسلامی ایران امروز شنبه با «بشار اسد» رئیس‌جمهور سوریه دیدار کرد.
- وزیر خارجه اسبق جمهوری اسلامی ایران با رئیس‌جمهور سوریه در دمشق دیدار کرد.

اگرچه هریک از روش‌های اول و دوم به خوبی عمل می‌کنند امّا تضمین نمی‌کنند که تمامی موارد دگربیان را بیابند؛ زیرا که با کاهش دامنه جستجو عملاً بخش عمده‌ای از عبارات مقایسه نمی‌شوند. این موضوع ایرادی برای روش‌های ارائه شده نخواهد بود؛ زیرا که نیازی به پیدا کردن تمامی عبارات و جملات دگربیان نیست و هدف افزایش دقت و سرعت الگوریتم استخراج خودکار، جهت ساخت پیکره متنی است.

برای بررسی عملکرد هر یک از دو روش در خارج از محدوده تعیین شده، تعداد ۱۰۰ نمونه از روش اول و ۱۰۰ نمونه از روش دوم استخراج می‌شوند تا عملکرد روش‌ها را در خارج از محدوده تعیین شده نیز بررسی نماییم. همچنین سعی شده است تا نمونه‌های تکراری از پیکره نهایی حذف شوند. با اعمال این روش‌ها بر روی داده‌های خام پیش‌پردازش شده، تعداد ۸۵۶ جفت نمونه از روش اول و ۶۶۷ جفت نمونه از روش دوم استخراج شد که جهت انجام عملیات نشانه‌گذاری مورداستفاده قرار می‌گیرد.

۴- نشانه‌گذاری به روش جمع‌سپاری

نشانه‌گذاری داده‌ها و پیکره‌ها بخش بسیار مهمی از آماده‌سازی را در برمی‌گیرد که لازم است با دقت و توجهی ویژه انجام شود. به‌طور معمول نشانه‌گذاری به کمک افراد متخصصی انجام می‌گیرد که به‌طور کامل جهت نشانه‌دار کردن پیکره وقت صرف می‌کنند. این روش معمولاً وقت‌گیر بوده و هزینه بالایی نیز دارد. در کنار این روش، جمع‌سپاری به عنوان یک روش جایگزین و با در نظر گرفتن خرد جمعی مطرح می‌شود که نتیجه سریع‌تر و ارزان‌تری را به ارمغان می‌آورد و به‌خوبی مورد توجه قرار گرفته است [۱۸، ۱۹].

در روش‌های جمع‌سپاری، وظیفه تهیه پیکره یا نشانه‌گذاری آن به‌جمعی از افراد علاقه‌مند یا متخصص سپرده می‌شود. هر یک از این افراد می‌توانند به تعداد دلخواه از نمونه‌های پیکره را نشانه‌گذاری نمایند و الزامی وجود ندارد که تمامی نمونه‌ها توسط هر فرد نشانه‌گذاری شوند.

امروزه ابزارهای مختلفی مانند سایت Amazon mechanical turk یا Crowdcrafting وجود دارند که به منظور جمع‌سپاری به کار می‌روند. بسیاری از پیکره‌هایی زبان‌شناسی رایانشی نیز به کمک تکنیک‌های جمع‌سپاری تهیه شده‌اند و از مزایای آن‌ها استفاده کرده‌اند [۶، ۲۰، ۲۱]. با وجود اینکه در بسیاری از موارد، مشارکت افراد به صورت داوطلبانه صورت می‌گیرد، امّا جهت تشویق افراد به مشارکت در ساخت پیکره معمولاً جایزه یا امتیازاتی برای آن‌ها در نظر گرفته می‌شود یا نحوه انجام کار را به صورت بازی و سرگرمی در می‌آورند تا نظر پاسخ‌دهنده به سؤالات را به خود جلب نمایند. به عنوان مثال با استفاده از سرویس ارائه‌شده توسط Amazon mechanical turk، می‌توان اعتبار لازم برای خریدهای اینترنتی از سایت آمازون را به دست آورد. نمونه‌های فارسی این ابزارها نیز در سال‌های اخیر طراحی شده‌اند که به صورت برنامه‌های موبایلی در دسترس قرار دارند و معمولاً برای پاسخ‌گویی به‌نظرسنجی‌ها به کار می‌روند.

با توجه به مزایای ذکر شده، از روش جمع‌سپاری جهت نشانه‌گذاری داده‌های به دست آمده استفاده می‌شود. جهت انجام عملیات نشانه‌گذاری، بر خلاف روش‌های پیشین جمع‌سپاری که از طریق سایت‌های اینترنتی یا نرم‌افزارهای طراحی شده کار می‌کردند، از روش جدیدی استفاده می‌شود که برمبنای پیام‌رسان تلگرام[۱۳] است [۲۲]. تلگرام پیام‌رسانی است که به‌خوبی در میان کاربران فارسی زبان شناخته شده است تا جایی که بیش از ۲۰ میلیون کاربر ایرانی در آن فعالیت دارند [۲۳]. علاوه بر ویژگی پیام‌رسانی این نرم‌افزار، ویژگی‌های دیگری مانند تماس صوتی و روبات‌ها نیز به این برنامه افزوده شده است و نسخه‌های مختلفی از آن در رایانه‌های رومیزی، صفحات وب و تلفن‌های همراه عرضه شده است که دامنه استفاده از آن را گسترش داده است. در روش پیشنهادی از روبات تلگرامی جهت عملیات جمع‌سپاری استفاده می‌شود. مزیت استفاده از روبات تلگرام، سهولت استفاده و در دسترس بودن آن برای کاربران ایرانی به همراه واسط برنامه‌نویسی غنی‌ای است که در اختیار قرار می‌دهد. برخلاف روش استفاده از سایت‌ها و نرم‌افزارهای جانبی که نیازمند نصب یا عضویت هستند، استفاده از ربات تلگرام نیاز به هیچ پیش‌زمینه‌ای ندارد و تنها از طریق آدرس نام‌کاربری ربات در داخل نرم‌افزار پیام‌رسان، امکان‌پذیر است که موجب برتری این روش نسبت به روش‌های دیگر می‌شود.

شکل ۲- سامانه نشانه‌گذاری به روش جمع‌سپاری به کمک روبات تلگرام

شکل ۲ نمایی از سامانه نشانه‌گذاری طراحی شده را نشان می‌دهد. برای پیاده‌سازی این سامانه از چارچوب برنامه‌نویسی بازمتنی، مبتنی بر زبان برنامه‌نویسی پایتون[۱۴] استفاده شده است. کاربر با استفاده از آدرس روبات می‌توان وارد آن شود. بعد از ورود کاربر، روبات جمع‌سپاری پیام ورود را برای کاربر ارسال می‌کند که شامل دکمه‌های تعاملی برای آغاز نشانه‌گذاری به همراه راهنمایی مختصری از نحوه کارکرد روبات و نکاتی در مورد نحوه نشانه‌گذاری است. با فشردن کلید شروع، یک جفت عبارت یا جمله به صورت تصادفی توسط روبات برای کاربر ارسال می‌گردد که باید قبلاً توسط کاربر نشانه‌گذاری نشده باشد. کاربر با توجه به راهنمایی‌هایی که توسط سامانه دریافت کرده و تشخیص خود می‌تواند به کمک دکمه‌های قرار داده شده، اقدام به نشانه‌گذاری نمونه کند. در صورت نیاز به ویرایش پاسخ می‌توان دکمه ویرایش قبلی را انتخاب کرد. همچنین امکانی برای نادیده‌گیری نمونه‌ها و پرش به نمونه بعدی وجود دارد و الزامی برای پاسخگویی به تمامی سؤالات نیست.

در این سامانه فرایند نشانه‌گذاری برای هر جفت نمونه، به کمک سه مقدار دگربیان (۱)، تقریباً دگربیان (۰) و نامرتبط (۱-) مشخص می‌شود که توسط کاربران مشخص می‌گردد. با در نظر گرفتن سه مقدار نشانه برای نمونه‌ها، عباراتی که نزدیک به دگربیان‌ها هستند نیز نشانه‌گذاری می‌شوند. عملیات تعریف و ارزیابی نشانه‌ها مشابه با عملیات ساخت پیکره دگربیان در زبان روسی در نظر گرفته شده است [۸].

یکی از مواردی که در روش‌های جمع‌سپاری در نظر گرفته می‌شود، جلوگیری از وارد شدن پاسخ‌های اتفاقی و نامربوطی است که به وسیله بعضی از کاربران در سامانه وارد می‌شود [۲۰]. در این موارد معمولاً با طراحی پرسش‌هایی که پاسخ بسیار ساده‌ای دارند، هوشیاری فرد پاسخ‌دهنده موردسنجش قرار می‌گیرد. از همین‌رو پرسش‌های در نظر گرفته شده شامل چنین عباراتی است که پاسخ واضحی را در بردارند. به عنوان مثال جفت عبارات کاملاً یکسانی در میان پرسش‌ها قرار گرفته‌اند تا با بررسی آن‌ها، پاسخ‌های اتفاقی شناسایی گردد. این پرسش‌ها با نام پرسش‌های ارزیابی مشخص می‌شوند و جنبه ارزیابی فرد پاسخ‌دهنده را دارند که در نسخه نهایی پیکره لحاظ نخواهند شد.

در فرآیند نشانه‌گذاری، هر یک از نمونه‌ها حداقل توسط ۳ کاربر نشانه‌گذاری می‌شوند. برای هریک از این افراد نتیجه سؤال ارزیابی، بررسی می‌شود و در صورتی که پاسخ آن اشتباه باشد تمامی داوری‌های صورت گرفته توسط آن فرد به دلیل نداشتن دقت کافی، در نظر گرفته نمی‌شوند. در پایان نشانه نهایی هر نمونه با محاسبه مقدار میانه تمامی داوری‌های صورت گرفته برای آن نمونه صورت می‌گیرد که یکی از مقادیر {۱، ۰٫۵، ۰، ۰٫۵-، ۱-} است. در صورتی که مقدار میانه برابر ۰٫۵ یا ۰٫۵- باشد مقدار صحیح قبلی؛ یعنی به ترتیب ۰ و ۱- به جای آن در نظر گرفته می‌شود. برای محاسبه برچسب نهایی هر نمونه روش‌های مختلفی وجود دارد که محاسبه مقدار میانه، یکی از این روش‌ها است که کارایی آن در ساخت پیکره به کمک جمع‌سپاری نشان داده شده است و در اینجا هم مورداستفاده قرار می‌گیرد.

جدول ۱- اطلاعات آماری نشانه‌گذاری‌های انجام‌شده توسط کاربران

متغیر آماری	مقدار
میانگین تعداد داوری‌های کاربران	۱۴۵٫۷۷
کمینه تعداد داوری‌ها	۱
بیشینه تعداد داوری‌ها	۱٫۲۹۵
میانگین امتیازات داوری	۰٫۸۸
تعداد داوران	۴۴

جدول ۱ اطلاعات آماری در خصوص تمامی داوری‌های جمع‌آوری شده توسط سامانه جمع‌سپاری را نشان می‌دهد. در مجموع ۴۴ کاربر جهت ساخت این پیکره همکاری کرده‌اند که حداقل تعداد داوری‌های انجام شده توسط کاربران، تعداد ۵ نمونه و حداکثر تعداد ۱٫۲۹۵ نمونه بوده است. به صورت میانگین هر کاربر تعداد ۱۴۵ داوری را انجام داده است. میانگین امتیازات داوری داده شده به نمونه‌ها هم مقدار ۰٫۸۸ را دارا است که نشان‌دهنده تمایل بیشتر کاربران برای نشانه‌گذاری نمونه‌ها با برچسب دگربیان است.

۵- ساختار پیشنهادی برای پیکره

در این قسمت ساختار پیشنهادی پیکره عبارات و جملات دگربیان را ارائه می‌دهیم. نمونه‌های استخراج شده بعد از نشانه‌گذاری، جهت ساخت پیکره نهایی استفاده می‌شوند. به منظور اعمال ویژگی قابل حمل بودن در پیکره، از قالب XML[۱۵] استفاده شده است تا سازگاری بهتری با ابزارهای مختلف برنامه‌نویسی داشته باشد. پیکره در دو فایل XML مجزا قرار گرفته است که شامل اطلاعات مربوط به جفت عبارات استخراج شده و داوری‌های انجام شده است.

در فایل مربوط به نمونه‌ها، شناسه هر نمونه توسط عنصر PairId مشخص شده است. عناصر Sentence1 و Sentence2 حاوی مقدار اصلی هر یک از عبارت‌های نمونه هستند. همچنین منبع هریک از دو عبارت نیز با استفاده از تگ‌های NewsSource و NewsId مشخص شده است که با استفاده از آن‌ها قادر هستیم تا به اصل خبرها دسترسی پیدا کنیم. مقدار MethodType روشی که این نمونه را استخراج کرده است نشان می‌دهد که یکی از دو مقدار method1 یا method2 را دارا است. عنصر Judge نیز داوری نهایی به دست آمده است که به صورت عددی ذخیره می‌شود.

به عنوان مثالی از ساختار فایل XML مربوط به نمونه‌ها خواهیم داشت:

```
<?xml version="1.0"?>
<PairCorpus>
<Pair>
<PairId>23466</PairId>
<Sentence1>
```

مصرف زیاد نمک به کبد نیز آسیب می‌زند.
</Sentence1>
<Sentence2>
تأثیر نمک بر فعالت کبد
</Sentence2>
<NewsSource1>asrIran</NewsSource1>
<NewsSource2>tabnak</NewsSource2>
<NewsId1>458654</NewsId1>
<NewsId2>566654</NewsId2>
<MethodType>method1</MethodType>
<Judge>1</ Judge >
</Pair>
</PairCorpus>

فایل XML مربوط به داوری‌ها هم به صورت زیر طراحی شده است:

```
<?xml version="1.0"?>
<JudgeCorpus>
<Pair>
<UserId>455315</UserId>
<PairId>23466</PairId>
<JudgeValue>1</JudgeValue>
</Judge>
</JudgeCorpus>
```

هر داوری با استفاده از سه عنصر UserId به عنوان شناسه داور، PairId به عنوان شناسه نمونه داوری شده و JudgeValue به عنوان مقدار داوری، مشخص می‌شود.

پیکره به دست آمده، به عنوان اولین داده تخصصی در رابطه با عبارات و جملات دگربیان، در سایت گروه پردازش زبان طبیعی دانشگاه گیلان به نشانی http://nlp.guilan.ac.ir در دسترس عموم قرار گرفته است.

۶- ارزیابی

داوری‌های به دست آمده، شامل ۶,۴۸۵ مورد است که از میان ۴۴ کاربر مختلف جمع‌آوری شده است. جدول ۲ نتایج نهایی به دست آمده از نشانه‌گذاری نمونه‌ها را به تفکیک روش اول و دوم نشان می‌دهد. درصد تکرار هر یک از رده‌ها در روش‌ها، در داخل پرانتز قرار داده شده است. به عنوان مثال، ۳۸ درصد از مجموع ۸۵۶ نمونه استخراج شده در روش اول را نمونه‌هایی با رده دگربیان تشکیل می‌دهند. با توجه به جدول، تعداد نمونه‌هایی که از طریق روش اول استخراج شده‌اند، نسبت به روش دوم، شامل تعداد بیشتر نمونه‌های دگربیان است. بخش اعظم نمونه‌های به دست آمده از روش دوم نیز شامل عبارات تقریباً دگربیان هستند که نشان‌دهنده آن است که بیشتر نمونه‌های استخراجی از این روش را جفت‌هایی تشکیل می‌دهند که معانی آن‌ها با اندکی تفاوت بیان شده است. هر یک از رده‌ها نیز شامل نمونه‌هایی در پیکره نهایی به دست آمده است.

جدول ۲- نتایج بدست آمده از نشانه‌گذاری نمونه‌ها

رده	تعداد نمونه‌های روش اول (%)	تعداد نمونه‌های روش دوم (%)
دگربیان	۳۲۸ (۳۸)	۲۰۷ (۳۱)
تقریباً دگربیان	۲۵۷ (۳۰)	۲۹۸ (۴۴)
نامرتبط	۲۷۱ (۳۱)	۱۶۲ (۲۴)
مجموع	۸۵۶	۶۶۷

مقادیر آستانه مشخص شده برای هر کدام از روش‌های اول و دوم به صورت تجربی به دست آمدند. برای ارزیابی درستی کارکرد این مقادیر، عملکرد روش‌ها را در داخل و خارج محدوده آستانه تعیین شده برای آن‌ها بررسی می‌کنیم. جدول ۳ و ۴ عملکرد هر کدام از این روش‌ها را در داخل و خارج محدوده تعیین شده مشخص می‌کند. از مجموع ۱۰۰ نمونه گنجانده شده در روش اول ۸۳ نمونه از آن‌ها در رده نامرتبط قرار گرفته‌اند و ۱۷ نمونه باقی‌مانده نیز به صورت دگربیان یا تقریباً دگربیان هستند. در روش دوم نیز ۷۸ نمونه از ۱۰۰ نمونه نامرتبط تشخیص داده شده‌اند. این مقادیر نشان می‌دهند که در هر دو این روش‌ها آستانه انتخاب شده، به درستی عمل کرده و با خروج از این محدوده تعداد مقادیر نامرتبط افزایش می‌یابد.

جدول ۳- ارزیابی حد آستانه تعیین شده در روش اول

رده	داخل محدوده (%)	خارج محدوده (%)	مجموع (%)
دگربیان	۳۲۲ (۴۲)	۶ (۶)	۳۲۸ (۳۸)
تقریباً دگربیان	۲۴۶ (۳۲)	۱۱ (۱۱)	۲۵۷ (۳۰)
نامرتبط	۱۸۸ (۲۴)	۸۳ (۸۲)	۲۷۱ (۳۱)
مجموع	۷۵۶	۱۰۰	۸۵۶

جدول ۴- ارزیابی حد آستانه تعیین شده در روش دوم

رده	داخل محدوده (%)	خارج محدوده (%)	مجموع (%)
دگربیان	۲۰۴ (۳۵)	۳ (۳)	۲۰۷ (۳۱)
تقریباً دگربیان	۲۷۹ (۴۹)	۱۹ (۱۹)	۲۹۸ (۴۴)
نامرتبط	۸۴ (۱۴)	۷۸ (۷۸)	۱۶۲ (۲۴)
مجموع	۵۶۷	۱۰۰	۶۶۷

۷- نتیجه‌گیری و کارهای آتی

در این مقاله توسعه پیکره جملات و عبارات دگربیان به تفصیل شرح داده شد. داده خام مورداستفاده شامل خبرهای بیش از ۱۱ خبرگزاری است که در مدت زمان دو ماه جمع‌آوری و سپس پیش‌پردازش شده است. دو روش به منظور استخراج بی‌ناظر عبارات و جملات دگربیان ارائه شد که در آن‌ها از معیار جاکارد استفاده شده است. در روش اول عبارات و جملات کوتاه دگربیان با بررسی عناوین موجود در خبرهای موجود در بازه زمانی یک روزه به دست آمدند. در روش دوم، متن خبرهایی که عناوین آن‌ها در روش اول استخراج شده بودند به عنوان خبرهایی هم موضوع انتخاب شدند و معیار جاکارد مجدداً بر روی تک‌تک جملات آن‌ها محاسبه شد و جملات دگربیان جدیدی استخراج گردید. سپس کل نمونه‌های استخراج شده، به کمک ابزار جمع‌سپاری جدیدی که تحت پیام‌رسان تلگرام پیاده‌سازی شده بود، نشانه‌گذاری شدند. فایل XML پیکره تولید شده به دلیل قابلیت حمل این نوع از فایل‌ها به راحتی توسط برنامه‌نویسان مختلف قابل استفاده است و نتیجه ارزیابی پیکره هم نشان می‌دهد که آستانه انتخاب شده به درستی کار می‌کند و نمونه‌هایی از هر سه رده در پیکره موجود است.

پیکره دگربیان به عنوان تنها پیکره فارسی تخصصی عبارات و جملات دگربیان، شامل ۱,۵۲۳ عدد نمونه است که از میان آن‌ها ۵۳۵ عدد به صورت دگربیان، ۵۵۵ عدد تقریباً دگربیان و ۴۳۳ عدد هم نامرتبط هستند. این پیکره از بسیاری از لحاظ قابل قیاس با نمونه‌های موجود در زبان‌های دیگر است و تلاش شده است تا ایرادات و ضعف‌های موجود در سایر پیکره‌ها در این پیکره وجود نداشته باشد. از ویژگی‌های مهم این پیکره وجود نمونه‌های با اندازه متنی کوتاه و بلند در کنار هم است. همچنین نمونه‌های موجود در آن محدود به جملات نیستند و با روشی که ارائه شده است عبارات دگربیان نیز استخراج شده‌اند. علاوه بر همه این ویژگی‌ها می‌توان به وجود سه رده برای نشانه‌گذاری نمونه‌ها اشاره کرد که

برخلاف بسیاری از پیکره‌های مشابه خارجی، نشانه‌گذاری‌ها برای نمونه‌های تقریباً دگربیان یا مشابه هم نیز انجام شده است.

نسخه اول پیکره به صورت عمومی در دسترس قرار گرفته است. نسخه‌های آتی پیکره نیز با هدف ادامه روند استخراج و نشانه‌گذاری نمونه‌ها ارائه خواهند شد. هدف ما این است که در دراز مدت، علاوه بر غنی‌سازی پیکره موجود، منابع داده‌ای بیشتری مانند ترجمه‌های موازی را نیز که به صورت ذاتی امکان وجود عبارات دگربیان در آن‌ها وجود دارد، جهت استخراج نمونه‌های دگربیان، به کار ببریم.

سپاسگزاری

پیکره حاضر نتیجه مشارکت کاربرانی است که نمونه‌های دگربیان را در سامانه جمع‌سپاری، نشانه‌گذاری کرده‌اند. نویسندگان این مقاله مایل هستند تا از تمامی افرادی که ما را در تهیه این پیکره یاری کردند، سپاسگزاری نمایند.

مراجع

[1] Y. Ji, and J. Eisenstein, "Discriminative improvements to distributional sentence similarity," in *EMNLP*, pp. 891–896, 2013.

[2] R. Bhagat, and E. Hovy, "What is a paraphrase?," *Computational Linguistics*, vol. 39, no. 3, pp. 463–472, 2013.

[3] B. Dolan, C. Quirk, and C. Brockett, "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources," in *20th international conference on Computational Linguistics*, p. 350, Association for Computational Linguistics, 2004.

[4] A. Eyecioglu, and B. Keller, "Asobek: Twitter paraphrase identification with simple overlap features and svms," *in SemEval*, 2015.

[5] W. Xu, A. Ritter, C. Callison-Burch, W. B. Dolan, and Y. Ji, "Extracting lexically divergent paraphrases from twitter," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 435–448, 2014.

[6] E. Pronoza, E. Yagunova, and A. Pronoza, "Construction of a russian paraphrase corpus: unsupervised paraphrase extraction," in *Information Retrieval*, pp. 146–157, Springer, 2016.

[7] P. M. McCarthy, and D. S. McNamara, "The user-language paraphrase corpus," *Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches: Issues and Approaches*, p. 73, 2011.

[8] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, "Ppdb: The paraphrase database," in *HLT-NAACL*, pp. 758–764, 2013.

[9] K. Khoshnavataher, V. Zarrabi, S. Mohtaj, and H. Asghari, "Developing monolingual persian corpus for extrinsic plagiarism detection using artificial obfuscation," in CLEF (Working Notes), 2015.

[۱۰] دبیرخانه شورای عالی اطلاع‌رسانی، "پیکرهٔ موازی انگلیسی- فارسی میزان،" http://dadegan.ir/catalog/mizan، ۱۳۹۲.

[11] T. Mosavi Miangah, "Constructing a large-scale englishpersian parallel corpus," *Meta*, vol. 54, no. 1, pp. 181–188, 2009.

[12] C. Boonthum, "istart: Paraphrase recognition," in *ACL 2004 workshop on Student research*, p. 55, Association for Computational Linguistics, 2004.

[13] V. Rus, R. Banjade, and M. C. Lintean, "On paraphrase identification corpora," in *LREC*, pp. 2422–2429, Citeseer, 2014.

[14] W. Xu, A. Ritter, and R. Grishman, "Gathering and generating paraphrases from twitter with application to normalization," in *the Sixth Workshop on Building and Using Comparable Corpora*, pp. 121–128, Citeseer, 2013.

[15] S. Wubben, A. Van Den Bosch, E. Krahmer, and E. Marsi, "Clustering and matching headlines for automatic paraphrase acquisition," in *the 12th European Workshop on Natural Language Generation*, pp.122–125, Association for Computational Linguistics, 2009.

[۱۶] سبحه، "هضم؛ پردازش زبان فارسی در پایتون،" http://www.sobhe.ir/hazm/، ۱۳۹۶.

[۱۷] دبیرخانه شورای عالی اطلاع‌رسانی، "ویراست یار؛ نرم‌افزار تخصصی ویرایش،" http://www.virastyar.ir، ۱۳۹۶.

[18] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines," in *LREC*, pp.859–866, 2014.

[19] M. Sabou, K. Bontcheva, and A. Scharl, "Crowdsourcing research opportunities: lessons from natural language processing," in *the 12th International Conference on Knowledge Management and Knowledge Technologies*, p. 17, ACM, 2012.

[20] S. M. Mohammad, B. J. Dorr, G. Hirst, and P. D. Turney, "Computing lexical contrast," *Computational Linguistics*, vol. 39, no. 3, pp. 555–590, 2013.

[21] E. Filatova, "Irony and sarcasm: Corpus generation and analysis using crowdsourcing," in *LREC*, pp. 392–398, 2012.

[22] Telegram, "Telegram bot api," https://core.telegram.org/bots/api, 2016.

[۲۳] خبرآنلاین، "نیمی از کاربران تلگرام ایرانی شدند،" http://www.entekhab.ir/fa/news/263281، ۱۳۹۵.

رضا معانی‌جو فارغ‌التحصیل سال ۱۳۹۲ کارشناسی رشته مهندسی کامپیوتر، گرایش نرم‌افزار است. از زمینه‌های مورد علاقه ایشان عبارتند از یادگیری ماشین، پردازش زبان‌های طبیعی، روباتیک، داده‌کاوی و یادگیری عمیق.
آدرس پست‌الکترونیکی ایشان عبارت است از:
rezamaanijou@msc.guilan.ac.ir

سید ابوالقاسم میرروشندل فارغ‌التحصیل از دانشکده فنی دانشگاه تهران در مقطع کارشناسی در رشته مهندسی کامپیوتر با گرایش نرم‌افزار و دانشگاه صنعتی شریف در مقاطع کارشناسی‌ارشد و دکتری در رشته مهندسی کامپیوتر گرایش هوش مصنوعی. از سال ۱۳۹۱ عضو هیات علمی گروه مهندسی کامپیوتر دانشگاه گیلان بوده و زمینه‌های مورد علاقه ایشان پردازش زبان‌های طبیعی، داده‌کاوی، یادگیری ماشینی و پردازش تصویر هستند.
آدرس پست‌الکترونیکی ایشان عبارت است از:
mirroshandel@guilan.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۶/۰۳/۰۷
تاریخ اصلاح: ۱۳۹۶/۰۴/۲۸
تاریخ قبول شدن: ۱۳۹۶/۰۵/۱۵
نویسنده مرتبط: دکتر سید ابوالقاسم میرروشندل، دانشکده فنی، دانشگاه گیلان، رشت، ایران.

---

[1]Paraphrase
[2]Information retrieval
[3]Text summarization
[4]Plagiarism detection
[5]Edit distance
[6]Twitter
[7]Crowdsourcing
[8]Crawler
[9]Stop words
[10]Term Frequency–Inverse Document Frequency
[11]Jaccard
[12]Wordnet
[13]Telegram
[14]Python
[15]Extensible Markup Language