اولین همایش ملی جویشگر بومی
۲۷ و ۲۸ بهمن ماه سال ۱۳۹۴
پژوهشگاه ارتباطات و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)
وزارت ارتباطات و فناوری اطلاعات
ITRC

سامانه پایه مرجع‌یابی گروه‌های اسمی در زبان فارسی با استفاده از قوانین ساده

شهره طباطبایی سیفی[۱]، یاسر شکفته[۲]

[۱] کارشناسی ارشد زبان شناسی رایانشی، پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی، تهران
tabatabaee-sh@rcdat.ir

[۲] استادیار پژوهشی، پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی، تهران
shekofteh@rcdat.ir

چکیده

مرجع‌یابی گروه‌های اسمی یکی از میان‌ابزارهای پردازش زبان طبیعی است که در بسیاری از سامانه‌های پردازش متن و زبان به کار گرفته می‌شود. در این مقاله چگونگی ساخت یک سامانه end-to-end مرجع‌یابی برای زبان فارسی تشریح شده است. این سامانه با استفاده از قوانین به نسبت ساده‌ای که بر روی مجموعه‌ای از مشخصه-مقدار ها اعمال می‌شوند پیاده‌سازی شده است. این پژوهش برای اولین بار گروه‌های اسمی را به عنوان بلاک‌های پایه در نظر گرفته است و از این نظر اولین نمونه در نوع خود در زبان فارسی به شمار می‌رود. استخراج مشخصه‌ها در این سامانه با روش‌های ساده پیکره محور انجام شده است ولی معماری پیمانه‌ای این سامانه به ما اجازه می‌دهد که هر کدام از ابزارهای زیرساختی را به راحتی با نمونه جدیدتر آن جایگزین کنیم اما ایراد این معماری این است که خطاهای رخ‌داده در مراحل اولیه در کل سامانه منتشر می‌شود و در دقت نهایی تاثیر به سزایی دارد. سامانه حاضر بر روی بخش آزمون پیکره اوپسالا آزمایش شده است و بر اساس معیار CONLL امتیاز ۴۸,۳۳ را به دست آورده است که می‌تواند به عنوان یک سامانه پایه برای مقایسه با سامانه‌های مشابه در نظر گرفته شود. همچنین با بهبود ابزارهای پیش‌پردازش استفاده شده در این سامانه این امتیاز ارتقای قابل ملاحظه‌ای پیدا می‌کند که در کارهای پیش رو انجام خواهد شد.

کلمات کلیدی

مرجع‌یابی، گروه اسمی، ابزارهای پیش پردازش متن، قوانین مرجع‌یابی

۱- مقدمه

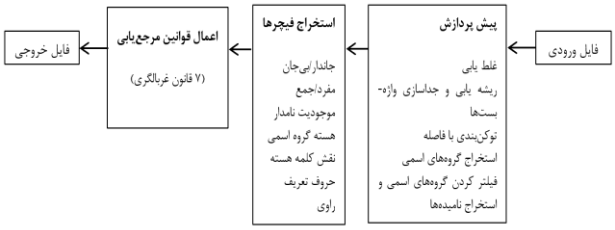
مساله مرجع‌یابی و یا به عبارت دیگر یافتن کلیه عباراتی که به یک موجودیت دلالت دارند، یک بخش عملیاتی مهمی در مسائلی مانند خلاصه‌سازی خودکار، پرسش و پاسخ خودکار و استخراج اطلاعات به شمار می‌رود.

سه روش عمده در مرجع‌یابی وجود دارد. در روش اول از قوانین قطعی برای یافتن عبارات هم‌مرجع استفاده می‌شود. رویکرد دوم استفاده از پیکره برچسب خورده و روش‌های یادگیری ماشین با ناظر و سومین رویکرد شامل استفاده از روش‌های بدون ناظر ولی مبتنی بر پیکره است. از آن جایی که در فارسی پیکره برچسب‌خورده قابل ملاحظه‌ای در حوزه مرجع‌یابی گروه-های اسمی موجود نیست باید از روش‌های بدون ناظر و یا مبتنی بر قوانین استفاده کنیم. روش‌های قانون محور در زبان انگلیسی نتایج خوبی را نشان داده‌اند به طوری‌که در رقابت‌های CONLL2012 [1] تعداد قابل توجهی از سامانه‌های شرکت داده شده در این همایش از روش‌های مبتنی بر قانون استفاده کرده‌اند.

مجموعه دلایل بالا ما را به این سمت سوق داد که یک روش مبتنی بر قانون ساده را پیاده-سازی کنیم. هدف اصلی این پژوهش تامین کلیه ابزارهای پیش‌پردازش مورد نیاز چنین سامانه-ای بوده است زیرا بسیاری از این ابزارها در زبان فارسی موجود نیست و یا دقت بسیار پایینی دارد. همچنین این سامانه به شکل end-to-end کار می‌کند یعنی فرض بر این است که باید مجموعه گروه‌های اسمی که می‌توانند در مرجع‌یابی شرکت کنند نیز توسط خود سامانه استخراج شود و از قبل به آن داده نمی‌شود. ساختار ادامه مقاله به این شرح است؛ در بخش دوم مروری بر کارهای مشابه انجام شده است. بخش سوم مقاله سامانه مرجع‌یابی پیشنهادی را توضیح می‌دهد. در بخش چهارم قوانین طراحی شده برای سامانه ارائه شده است و بخش پنجم به ارزیابی و بررسی نتایج می‌پردازد. در انتها نیز جمع‌بندی و نتیجه‌گیری مقاله آورده شده است.

۲- مروری بر کارهای مشابه

سامانه‌های زیادی در زبان انگلیسی برای مرجع‌یابی ارائه شده است. به خصوص در سال ۲۰۱۲ که کنفرانس CONLL رقابت اصلی خود را ارائه سامانه‌های مرجع‌یابی قرار داد و به این وسیله گروه‌های تحقیقاتی زیادی تمرکز خود را بر این عملیات قرار دادند. در ادامه مروری کلی بر روش‌های استفاده شده در سامانه‌های مرجع‌یابی در زبان انگلیسی و چند نمونه موجود در زبان فارسی انجام خواهد شد.



شکل ۱ : معماری سامانه مرجع‌یابی پیشنهادی

۲-۱- روش‌های یادگیری ماشین

روش‌های مبتنی بر کلاسه بندی ابتدا نمونه های مثبت شامل «ضمیر، مرجع درست» و نمونه های منفی شامل «ضمیر، مرجع اشتباه» را تشکیل می‌دهند و سپس از آن‌ها برای آموزش یک کلاسه‌بند استفاده می‌کنند و معیار f ۶۰٪ دست پیدا می‌کند. [3]

در مقاله [4] یک روش خوشه بندی استفاده شده است که در آن هر عبارت اسمی تبدیل به یک بردار مشخصه-مقدار می‌شود و سپس خوشه بندی می‌شود. مقاله [13] یک رویکرد رتبه بندی با ناظر ارائه می‌کند و به جای این که دو به دو کاندیداها را مقایسه کند، همه کاندیداها را همزمان بررسی می‌کند. گراف، مدل درجه اول احتمالاتی و داده‌کاوی از دیگر روش‌هایی هستند که برای مرجع‌یابی استفاده شده‌اند. سادات موسوی و قاسم‌ثانی در [2] بخشی از پیکره بیجن‌خان را برچسب‌زنی کرده‌اند که PCAC نام دارد. در این پیکره نزدیکترین مرجع هر ضمیر مشخص شده است و ۳۰ سند و ۲۰۰۶ ضمیر موجود در آن برچسب خورده‌اند. آن‌ها از۲۵ مشخصه استفاده کرده‌اند و۴ الگوریتم درخت تصمیم، SVM، پرسپترون و ماکسیمم آنتروپی را به کار گرفته‌اند. بهترین F-measure به دست آمده ۴۴٫۷۵٪ و با الگوریتم C4.5 بوده است.

۲-۲- روش‌های مبتنی بر قانون

ایده استفاده از مجموعه‌ای از قوانین، که از قوانین با دقت بالا شروع می‌شود و با اضافه کردن قانون‌های جدید سعی در بالا بردن بازیابی دارد، اولین بار توسط Baldwine [6] ارائه شد. او پیشنهاد داده بود که از ۷ قانون به عنوان فیلترهایی برای استخراج روابط هم‌مرجعی استفاده شود. ایده استفاده از قوانین بعدتر توسط [7] و تعداد زیادی دیگر از محققان به کار گرفته شد. اما تا قبل از [8] هیچ کدام از سامانه‌های مرجع‌یابی مبتنی بر قانون موفق به رسیدن به بازدهی‌های گزارش شده توسط سامانه‌های مبتنی بر یادگیری ماشین نشدند. اما در رویکرد ارائه شده توسط حقیقی در [8] برای نخستین بار بازدهی یک سامانه مبتنی بر قوانین در حد و اندازه‌های سامانه‌های یادگیری ماشین با ناظر و حتی گاهی جلوتر از آنها بود.

پس از حقیقی، محققان دانشگاه استنفورد در طی چند مقاله که در سال‌های ۲۰۱۰، ۲۰۱۱ و ۲۰۱۳ [9] [10] [11] منتشر کردند، سامانه مبتنی بر قانونی را ارئه کردند که از کلیه سامانه‌های ارائه شده تا آن زمان پاسخ بهتری داشت. قوانین استفاده شده در این سامانه سلسله‌مراتبی نسبتا ساده و مستقل از زبان انگلیسی هستند و به دلیل همین خصوصیت اقبال زیادی به آن شد به طوری که در کنفرانس CONLL2012 که در آن مساله مرجع‌یابی برای سه زبان انگلیسی، عربی و چینی مورد نظر بود، بیش از ۴۰ درصد سامانه‌های شرکت کرده از رویکرد ارایه شده در استنفورد استفاده کردند.

فلاحی، شمس فرد مقاله‌ای در سال ۲۰۱۱ ارائه کرده‌اند که در آن با استفاده از قوانین عملیات یافتن مرجع را انجام داده‌اند. [12] آن‌ها در فاز پیش‌پردازش عملیات تعیین مرز کلمات، ریشه یابی، برچسب‌زنی مقوله نحوی (POS) را انجام می‌دهند. برای تست ۱۰۰ صفحه از ۵ وبلاگ که به شکل دستی برچسب خورده است استفاده کرده‌اند و به دقت ۹۵ و بازیابی ۹۰ دست یافته‌اند.

۳- سامانه مرجع‌یابی پیاده‌سازی شده

سامانه جامع پیاده‌سازی شده در زبان فارسی دارای خصوصیات منحصر به فردی است که به نوعی اولین سامانه مرجع‌یابی فارسی در نوع خود به شمار می‌رود. در این سامانه گروه‌های اسمی عناصر پایه‌ای هستند که در مرجع‌یابی شرکت می‌کنند در حالی که در سایر مقالات ارائه شده در این حوزه در زبان فارسی، فقط کلمات مورد توجه هستند. همچنین در سامانه حاضر زنجیره‌های هم‌مرجع استخراج می‌شوند که در این زنجیره‌ها هم ضمایر قرار دارند و هم گروه-های اسمی در حالیکه در سایر مقالات ارائه‌شده در این حوزه اغلب، فقط مرجع ضمایر تشخیص داده می‌شود.

اگرچه دقت ابزارهای پیش پردازشی زبان فارسی فاصله بسیار زیادی با نمونه‌های مشابه در زبان انگلیسی دارد ولی در این پژوهش تلاش شده است بر اساس استانداردهای موجود در این حوزه پیاده‌سازی صورت پذیرد. این رویکرد به ما کمک می‌کند که سامانه ارائه‌شده قابل قیاس با نمونه‌های موجود در زبان‌های دیگر باشد به خصوص مقالات ارائه‌شده در کنفرانس CONLL 2012 که نتایج سامانه‌ها را بر روی پیکره OntoNote2012 نشان می‌دهد. پیکره OnToNote 2012 در سه زبان انگلیسی، چینی و عربی تهیه شده است. این پیکره دارای برچسب‌های POS، درخت تجزیه سازه‌ای و زنجیره‌های هم‌مرجع است. معماری سامانه در شکل ۱ دیده می‌شود.

۳-۱-پیش پردازش

در این بخش عملیات زیر بر روی فایل ورودی انجام می‌شود.

- **غلط‌یابی**

غلط‌یاب سه مجموعه عملیات نرمال‌سازی متن و جداسازی جملات و اعمال غلط‌یاب ویراستیار را بر روی متن انجام می‌دهد. نرمال‌سازی با استفاده از مجموعه‌ای از قوانین اصلاح کاراکترها صورت می‌پذیرد. جداسازی جملات نیز با استفاده از مجموعه‌ای قوانین نگارشی انجام می‌شود. قسمت اصلی ماژول غلط‌یاب اعمال ویراستیار[1] بر روی متن ورودی است. ویراستیار با تنظیماتی که در آن هر غلط پیدا شده را با اولین کلمه درست در فهرست کلمات جایگزین می-کند اجرا می‌شود.

- **ریشه‌یابی**

زبان فارسی دارای واژه‌بست‌های ضمیری است که بدون فاصله به کلمه ماقبل خود می‌چسبد. مانند ضمیر ملکی «م» که به انتهای کتاب می‌چسبد و تشکیل کلمه ترکیبی «کتابم» را می-دهد. این ضمایر باید در عملیات مرجع‌یابی به طور جدا شرکت داده شوند و به همین دلیل باید بعد از غلط‌یابی، عملیات جداسازی این مجموعه واژه‌بست‌ها انجام شود. این جداسازی با استفاده از مجموعه‌ای از قوانین و برچسب مقوله نحوی کلمات صورت می‌پذیرد.

- **توکن‌بندی**

سومین مرحله عملیات پیش‌پردازش جداسازی کلمات با استفاده از فاصله است. خروجی این بخش هر توکن را در یک خط در فایل خروجی قرار می‌دهد. پایان هر جمله با یک خط خالی مشخص می‌شود.

- **استخراج گروه‌های اسمی**

اولین قدم در یک سامانه end-to-end جداسازی گروه‌های اسمی است. در سامانه‌های موجود در زبان‌های دیگر، این بخش با استفاده از ابزار تجزیه درختی سازه‌ای انجام می‌شود. به این ترتیب که گروه‌های اسمی موجود در درخت تجزیه به ترتیب اندیس‌گذاری می‌شوند. ممکن است برخی از گروه‌های اسمی تو در تو باشند که با استفاده از برخی قوانین ساختی تعدادی از آن‌ها حذف می‌شوند. اما در فارسی تجزیه‌گر سازه‌ای با دقت مناسب موجود نیست.

با توجه به شرایط فوق برای استخراج گروه‌های اسمی از یک تجزیه‌گر سطحی آموزش داده-شده بر روی پیکره گروه‌های نحوی استفاده شده است. این تجزیه‌گر سطحی توسط طباطبایی و حسین‌نژاد تولید شده و در [۱۵] خصوصیات آن توضیح داده شده است. بنا بر این مقاله کارآیی این تجزیه‌گر سطحی کمی بیش از ۹۰٪ در معیار F است.

- **استخراج نامیده‌ها[2]**

در بخش استخراج نامیده‌ها از میان گروه‌های اسمی آن دسته از گروه‌های اسمی که نمی‌توانند مرجع قرار بگیرند حذف می‌شوند. به عنوان مثال گروه اسمی «هر کشاورز» اگرچه یک گروه اسمی است ولی به دلیل وجود سور «هر» در ابتدای آن به هیچ موجودیتی اشاره نمی‌کند و نمی‌تواند نامیده باشد. لیست نامیده‌ها به مرحله استخراج مشخصه‌ها و اعمال قوانین مرجع‌یابی به عنوان ورودی تحویل داده می‌شود. اگر این بخش به اشتباه یک گروه اسمی صحیح را حذف کند بازیابی کل سامانه افت می‌کند ولی اگر برخی از گروه‌های اسمی به خطا به عنوان نامیده باقی بمانند شاید در مرحله قوانین مرجع‌یابی حذف شوند، به این شکل که توسط هیچ قانونی به عنوان عضوی از یک زنجیره در نظر گرفته نشوند. به همین دلیل ما از مجموعه قوانینی در این مرحله استفاده کردیم که مطمئن باشیم هیچ نامیده‌ای به اشتباه حذف نمی‌شود. اگر یک گروه اسمی که بالقوه می‌تواند نامیده باشد در این بخش استخراج نشود خطای آن در مراحل بعدی منتشر می‌شود.

۳-۲- استخراج مشخصه‌ها

پس از استخراج لیست نامیده‌ها یک بردار مشخصه-مقدار برای این مجموعه گروه اسمی تشکیل می‌شود. دقت سامانه نهایی به شدت به دقت این مشخصه-مقدارها بستگی دارد. این مشخصه‌ها عمدتا با استفاده از ابزارهای پردازشی پایه به دست می‌آیند ولی از آن جایی که بیشتر این ابزارها یا در زبان فارسی موجود نیستند و یا این که در دسترس ما نبودند، با روش-

های ساده کلیه این ابزارها ساخته شدند. معماری پیمانه‌ای (ماژولار) این سامانه به ما اجازه می‌دهد که هر وقت یکی از ابزارهای پایه با دقت مناسب تولید شد، به راحتی در جایگاه خود به جای ابزار ساده فعلی استفاده شود. برخی از مشخصه‌هایی که در این سامانه استفاده می‌شود شامل: نقش هسته نامیده، مفرد یا جمع ، جاندار یا بی‌جان ، موجودیت نامدار و ضمیر است.

تشخیص هسته با استفاده از ساخت درونی گروه اسمی صورت می‌پذیرد. نقش دستوری هسته توسط تجزیه‌گر وابستگی که روی داده‌های بانک درخت وابستگی دادگان آموزش داده شده است تعیین می‌شود. تشخیص حضور نامیده در بخش راوی در یک جمله با تعیین وجود «:» قبل یا بعد از نامیده انجام می‌شود. چگونگی استخراج سه مشخصه بی‌جان/جاندار، جمع/مفرد و موجودیت نامدار در ادامه آمده است.

- **تعیین جاندار یا بی‌جان بودن**

مشخصه جاندار یا بی‌جان بودن یکی از مهم‌ترین مشخصه‌هایی است که در تعیین مرجع ضمایر نقش مهمی ایفا می‌کند. در فارسی هیچ ابزار تشخیص جاندار یا بی‌جان در مورد اسامی وجود ندارد. پیکره درخت‌بانک وابستگی دادگان دارای دو مجموعه برچسب مقوله نحوی است. در مجموعه برچسب‌های ریزدانه که شامل ۴۰ برچسب است درمورد اسامی مشخصه جاندار-بی‌جان وجود دارد. با استفاده از این پیکره یک برچسب‌زن مقوله نحوی آموزش داده شد و جاندار یا بی‌جان بودن با آن تعیین شد. یک نامیده در صورتی جاندار /بی‌جان است که کلمه هسته در این نامیده جاندار/بی‌جان باشد.

- **جمع یا مفرد**

تشخیص جمع یا مفرد بودن یک نامیده، هم در تشخیص مرجع ضمیر موثر است و هم در ساخت زنجیره‌های گروه‌های هم‌مرجع. یک روش رایج در استخراج این مشخصه استفاده از مجموعه‌ای از قوانین است که کلمات جمع را مشخص می‌کند. البته از آن جایی که تعداد بسیار زیادی جمع مکسر در فارسی استفاده می‌شود باید به این قوانین استثنائات نیز با یک لغت‌نامه اضافه شود. ما در این پژوهش از این روش استفاده نکردیم زیرا معمولا لیست جمع‌های مکسر هیچ‌گاه به اندازه کافی کامل نیستند. در پیکره بیجن‌خان برای کلیه اسامی جمع یا مفرد بودنشان به عنوان برچسب‌های زیرمقوله‌ای تعیین شده است. با استفاده از یک برچسب‌زن مقوله نحوی که دارای ۱۰۵ برچسب است و برای کلیه اسامی جمع یا مفرد بودنشان را تعیین می‌کند، مشخصه جمع یا مفرد بودن استخراج شده است. تشخیص این که یک نامیده جمع است یا مفرد کمی پیچیده‌تر از بخش قبلی است. در این حالت هم باید کلمه هسته مورد توجه قرار بگیرد و هم ساخت‌های عطفی که هسته هر کدام از گروه‌ها در آن مفرد است ولی با یکدیگر یک مرجع برای ضمایر جمع می‌سازند.

- **موجودیت نامدار**

اگر یک موجودیت نامدار در جمله‌ای وجود داشته باشد احتمال این که مرجع یک ضمیر واقع شود بسیار بیشتر از یک اسم عام است. تشخیص موجودیت نامدار معمولا با یک برچسب‌زن NER انجام می‌شود. اگرچه هم‌اکنون یک برچسب‌زن NER در اختیار این گروه پژوهشی قرار دارد ولی در زمان پیاده‌سازی این سامانه هنوز این ابزار آماده نبود و به همین دلیل از برچسب مقوله نحوی کلمات استفاده کردیم. به این ترتیب که کلیه اسامی که اسم خاص تشخیص داده می‌شدند به عنوان موجودیت نامدار در نظر گرفته شدند.

۴- قوانین مرجع‌یابی

۷ قانون بر روی نامیده‌ها اجرا می‌شود که به این ترتیب گروه‌هایی که با یکدیگر هم‌مرجع هستند مشخص می‌شوند. این دسته از نامیده‌ها دارای شماره زنجیره یکسانی در انتهای عملیات خواهند بود.

قانون اول: اگر دو نامیده عینا یکی باشند هم‌مرجع می‌شوند.

قانون دوم: اگر دو نامیده دارای عبارات تعریف یکسانی باشند (این، همین، همان) و هسته گروه اسمی بعد از آن مشابه باشد با یکدیگر هم‌مرجع می‌شوند.

قانون سوم: اگر دو عبارت موجودیت باشند و دارای هسته مشابه هم باشند با یکدیگر هم‌مرجع می‌شوند.

قانون چهارم: مرجع ضمایر متکلم را در بخش راوی جملات مشخص می‌شود.

قانون پنجم: مرجع ضمایر سوم شخص مفرد (او، ش، اش، خود و خودش و...) از میان نامیده‌های قبلی که جاندار و مفرد باشند، انتخاب می‌شود.

قانون ششم: مرجع ضمایر سوم شخص جمع (آنها، آنان، ایشان و ...) از میان نامیده‌های قبلی که جاندار و جمع باشند انتخاب می‌شود.

معیار ارزیابی	Recall	Precision	F-measure
MUC	47.12	37.87	47.12
B-3	44	71	54.35
CEAFe	58.4	29.9	39.55
CEAFm	43.51	43.51	43.51
CONLL	-----	-----	**48.33**

جدول ۱: ارزیابی سامانه مرجع‌یابی پیشنهادی با نامیده های طلایی

قانون هفتم: مرجع ضمایر سوم شخص غیر جاندار (جمع یا مفرد) از میان نامیده‌هایی که مشخصات مشابه را داشته باشند انتخاب می‌شود.

۵- ارزیابی سامانه مرجع‌یابی

انتخاب معیار مناسب برای ارزشیابی سامانه‌هایی که عملیات مرجع‌یابی را به طور خودکار انجام می‌دهند همواره مورد توجه محققان این حوزه بوده است. در این راستا از سال ۱۹۹۵ تا کنون معیارهای مختلفی ارائه شده که هر کدام به نوعی تلاش کرده است کیفیت سامانه مرجع‌یابی را بهتر نمایش دهد. در این راستا به مرور معیارهای متفاوتی ارائه شده است که چون هرکدام به نوعی کیفیت سامانه ارائه شده را نشان می‌دهد در رقابت‌های مرتبط با مرجع‌یابی تقریبا همه این معیارها در نظر گرفته می‌شوند.

برای ارزیابی سامانه حاضر بخش آزمون پیکره اوپسالا [14] مورد به شکل دستی و با شیوه ارائه شده در برچسب‌زنی پیکره Ontonote2012 برچسب‌گذاری شد. پیکره اوپسالا یک درخت‌بانک وابستگی است که بر روی بخشی از پیکره بیجن‌خان ساخته شده است. این پیکره دارای ۶۰۰۰ جمله است و بخش آزمون آن که ۶۰۰ جمله و بیش از ۱۶۰۰۰ کلمه دارد برای برچسب‌زنی هم‌مرجع انتخاب شده است. ارزیابی بر اساس چهار معیار MUC، B3 و CEAF-e و CEAF-m صورت پذیرفته است که معیارهای استاندارد در ارزیابی سامانه‌های مرجع‌یابی هستند. برای این که امکان مقایسه راحت‌تر فراهم شود معیار CONLL نیز اضافه شده است که در حقیقت میانگین معیارهای اول تا سوم است. این معیار در رقابت‌های مرجع-یابی به طور استاندارد استفاده می‌شود.

جدول ۱ زیر نتایج به دست آمده از اعمال این سامانه بر روی پیکره آزمون را نشان می‌دهد. در این آزمایش از نامیده‌های طلایی استفاده شده است. یعنی بخش انتخاب نامیده‌ها را از عملیات حذف کردیم. جدول ۲ نتایج به دست آمده توسط بهترین سامانه‌های ارائه شده در رقابت‌های CONLL2012 را نشان می‌دهد. این نتایج نیز با استفاده از نامیده‌های طلایی به دست آمده

است. جدول ۳ نتایج بخش استخراج نامیده‌ها را در سامانه پیشنهادی نشان می‌دهد. باید توجه داشت که پایین بودن دقت در این بخش چندان مهم نیست زیرا ممکن است حجم زیادی گروه اسمی تولید شود که در هیچ زنجیره‌ای حضور ندارند ولی پایین بودن بازیابی تاثیر منفی زیادی در سامانه نهایی خواهد داشت. جدول ۴ نتایج ارزیابی سامانه پیشنهادی end-to-end را نشان می‌دهد. افت بسیار شدید نتایج حاصل از خطای بالا در بخش انتخاب نامیده‌ها است. این خطا نشان می‌دهد که عدم وجود یک تجزیه‌گر سازه‌ای کارا و مناسب و استفاده از تجزیه‌گر سطحی تا چه حد می‌تواند در پایین آمدن کارآیی سامانه موثر باشد. نکته قابل توجه این است که کارآیی تجزیه‌گر سطحی استفاده شده، به عنوان یک ابزار گروه‌بندی نحوی نسبتا قابل قبول است (معیار F این ابزار بیش از ۹۰٪ است) و اغلب گروه‌های اسمی که پیدا می‌کند صحیح هستند ولی مساله این است که معمولا گروه‌های اسمی دارای ساختار سلسله مراتبی هستند و یک نامیده ممکن است زیر مجموعه‌ای از یک گروه اسمی باشد. به عنوان مثال در عبارت «دیدار رییس جمهور با سران کشورهای اسلامی» گروه اسمی «رییس جمهور» باید به عنوان یک نامیده انتخاب شود در حالیکه خروجی تجزیه‌گر سطحی گروه اسمی بزرگتر یعنی «دیدار

Participant	CONLL SCORE		
	English	Chinese	Arabic
fernandez	69.35	66.36	**63.49**
chen	70.46	**77.77**	52.26
chang	**77.22**		

جدول ۲: بهترین نتایج به دست آمده در رقابت‌های CONLL 2012 با استفاده از نامیده های طلایی. [1]

رییس جمهور» را به عنوان خروجی تولید می‌کند که از نظر نحوی کاملا صحیح است ولی امتیاز منفی می‌گیرد.

انتخاب نامیده‌ها در سامانه پیشنهادی	Recall	Precision	F-measure
	59	10.94	18.46

جدول ۳: نتایج بخش انتخاب نامیده‌ها در سامانه پیشنهادی، پایین بودن recall تاثیر منفی به سزایی در نتیجه نهایی سامانه دارد

معیار ارزیابی	Recall	Precision	F-measure
MUC	10.29	12.26	11.19
B-3	18.69	8.88	12.04
CEAFe	2.61	43	4.93
CEAFm	29.41	5.44	9.19
CONLL	----	----	**9.38**

جدول ۴: نتایج ارزیابی سامانه پیشنهادی به صورت end-to-end و بدون استفاده از نامیده های طلایی

۶- تحلیل نتایج

عمده اشتباهات از خطاهای ایجاد شده در بردار مشخصه-مقدار می‌آیند. در روش‌های مبتنی بر قانون، دقت بردارهای مشخصه-مقدار نقش تعیین کننده دارند. به عنوان مثال تعیین جمع بودن یک گروه اسمی در روش حاضر با استفاده از یک برچسب‌زن مقوله نحوی انجام می‌شود. این رویکرد اگرچه پوشش نسبتا خوبی دارد ولی در تعیین گروه‌های اسمی جمعی که هسته آن‌ها مفرد است دچار خطا می‌شود. به عنوان مثال عبارت «هزار سرباز» به اشتباه گروه مفرد در نظر گرفته می‌شود. یکی دیگر از منابع بروز خطا اشتباه در تعیین جاندار/بی‌جان بودن یک گروه اسمی است به عنوان مثال ضمیر «او» فقط مرجع جاندار را قبول می‌کند اما در جمله «تخیل ایرانی این گونه مطالب را بی‌شاخ و برگ نمی‌گذارد. او دلخواه خود را بر آن می‌افزاید.» چون عبارت «تخیل ایرانی» برچسب بی‌جان دارد نمی‌تواند مرجع «او» قرار بگیرد. دسته دیگری از خطاها به دلیل عدم تشخیص ارتباطات معنایی بین گروه‌های اسمی است.

۷- کارهای در دست انجام

بالا بردن دقت تک‌تک ابزارهای پیش‌پردازشی که منجر به استخراج بردار مشخصه-مقدار می‌شوند اولین قدمی است که باید در این سامانه پیاده‌سازی شود. مهم‌ترین بخش آن نیز بالابردن دقت انتخاب گروه‌های اسمی مناسب برای شرکت در مرجع‌یابی است. این عمل ممکن نخواهد بود مگر با در دست داشتن یک تجزیه‌گر سازه‌ای با دقت بالا که بتواند گروه‌های اسمی را به درستی استخراج کند. همچنین ابزارهای تشخیص موجودیت نامدار، تشخیص جاندار/بی‌جان و تشخیص مفرد/جمع بودن گروه‌های اسمی باید هر کدام به صورت جداگانه مورد بررسی قرار بگیرند و به شکل ماژول‌های با دقت بالا تولید شوند. در این راستا پیکره موجودیت نامدار در این گروه پژوهشی تهیه شده است و پروژه تولید درخت‌بانک سازه‌ای نیز آغاز شده است.

۸- نتیجه‌گیری

مرجع‌یابی یکی از میان‌ابزارهایی است که در عملیات متنوعی از جمله استخراج اطلاعات، خلاصه‌سازی متون، ساخت هستان‌شناسی‌ها از روی متون و غیره کاربرد دارد. در این پژوهش یک سامانه مرجع‌یابی مبتنی بر قوانین ارائه شده است که در مقایسه با نمونه‌های مشابه در زبان فارسی دامنه گسترده‌تری را در بر می‌گیرد زیرا هم گروه‌های اسمی را به عنوان عناصر پایه در نظر می‌گیرد و هم زنجیره‌های هم‌مرجع را استخراج می‌کند در حالیکه در برخی پژوهش‌های مشابه فقط نزدیک‌ترین مرجع در نظر گرفته شده است. ارزیابی F این سامانه در حالتی که گروه‌های اسمی طلایی از قبل مشخص شده باشد ۴۸,۳۳ در مقیاس CONLL است که می‌تواند به عنوان یک سامانه مبنا در نظر گرفته شود. البته این سامانه در مقایسه با بهترین سامانه موجود در زبان عربی که ارزیابی F آن ۶۳,۴۹ در مقیاس CONLL است می‌تواند قابل رقابت در نظر گرفته شود و با اصلاحات پیشنهاد شده در بخش قبل، هم‌تراز با آن قرار بگیرد.

۹- مراجع

[1] Pradhan, Sameer, et al. "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes." Joint Conference on EMNLP and CoNLL-Shared Task. Association for Computational Linguistics, 2012.

[2] Nafiseh Sadat Moosavi and Gholamreza Ghassem-Sani (2009). A Ranking Approach to Persian Pronoun Resolution. Advances in Computational Linguistics. Research in Computing Science, 41, 169-180.

[3] Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: ACL, pp. 104--111 (2002b).

[4] Wagstaff, K.: Intelligent Clustering with Instance-Level Constraints. Ph.D. dissertation, Department of computer science and engineering, Cornell University 2002.

[5] Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., Roukos, S.: A mention-synchronous coreference resolution algorithm based on the bell tree. In: ACL, pp. 136--143 2004.

[6] Baldwin, Breck. CogNIAC: A Discourse Processing Engine. University of Pennsylvania Department of Computer and Information Sciences. Ph.D. thesis. 1995

[7] Haghighi, Aria, and Dan Klein. "Unsupervised coreference resolution in a nonparametric bayesian model." Annual meeting-Association for Computational Linguistics. Vol. 45. No. 1. 2007.

[8] Haghighi, Aria, and Dan Klein. "Simple coreference resolution with rich syntactic and semantic features." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. Association for Computational Linguistics, 2009.

[9] Raghunathan, Karthik, et al. "A multi-pass sieve for coreference resolution." Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010.

[10] Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.

[11] Lee, Heeyoung, et al. "Deterministic coreference resolution based on entity-centric, precision-ranked rules." Computational Linguistics 39.4 .2013: 885-916.

[12] Fallahi, Farshid, and Mehrnoush Shamsfard. "Recognizing Anaphora Reference in Persian Sentences." IJCSI .2011: 324.

[13] Denis, P., Baldridge, J.: A Ranking Approach to Pronoun Resolution. In: IJCAI, pp. 1588--1593 .2007

[14] Seraji, Mojgan, et al. "A Persian Treebank with Stanford Typed Dependencies." The 9th International Conference on Language Resources and Evaluation (LREC), 2014, 26-31 May, Reykjavik, Iceland. 2014.

[۱۵] طباطبایی سیفی، شهره، حسین‌نژاد، شادی، تعیین مرز و نوع عبارات نحوی با استفاده از پیکره تولید شده از بانک درختان وابستگی دادگان، سومین همایش زبان‌شناسی رایانشی، دانشگاه صنعتی شریف، ۱۳۹۳

زیرنویس‌ها

[۱] نرم‌افزار متن‌باز شورای عالی اطلاع‌رسانی

[۲] در این مقاله از معادل «نامیده» به جای mention استفاده می‌شود. نامیده به آن دسته از گروه‌های اسمی گفته می‌شود که می‌توانند به یک موجودیت اشاره کنند و مرجع قرار بگیرند.