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

بررسی تاثیر دقت برچسب اجزای کلام در کارایی سامانه شناسایی موجودیت‌های نامدار برای زبان فارسی

شادی حسین‌نژاد[۱] یاسر شکفته[۲]، طاهره امامی آزادی[۳]

[۱] گروه پردازش صوت و زبان طبیعی، پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی ، تهران،
Hosseinnejad@rcdat.ir

[۲] گروه پردازش صوت و زبان طبیعی، پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی ، تهران،
shekofteh@rcdat.ir

[۳] گروه پردازش صوت و زبان طبیعی، پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی ، تهران،
t.emami@rcdat.ir

چکیده

تشخیص موجودیت‌های نامدار یکی از مسائل پردازش زبان طبیعی است که هدف آن شناسایی موجودیت‌های نامدار موجود در یک متن و دسته‌بندی آنها در دسته‌های از پیش تعیین شده است. یکی از روش‌های تشخیص موجودیت نامدار، استفاده از پیکره برچسب‌گذاری شده و اعمال روش‌های یادگیری ماشین است. در این مقاله نحوه تولید یک سامانه تشخیص موجودیت نامدار مبتنی بر پیکره با استفاده از روش میدان‌های تصادفی شرطی شرح داده می‌شود. همچنین تاثیر دقت برچسب اجزا کلام (به عنوان یک ویژگی مورد استفاده) بر دقت سامانه موجودیت نامدار بررسی می‌گردد. سامانه ذکر شده با استفاده از پیکره فارسی اعلام که حاوی حدود ۲۵۰ هزار توکن است، تولید شده است. آموزش این سامانه با استفاده از برچسب‌های اجزای کلام دقیق برگرفته از پیکره متنی فارسی منتج به f-measure ۸۵ درصد شده است. این در حالی است که استفاده از برچسب های تولیده شده از یک ابزار برچسب‌گذاری اجزای کلام (برچسب‌گذاری غیر دقیق) در آموزش این سامانه منجر به دستیابی به f-measure ۸۴/۹ درصد شده است.

کلمات کلیدی

تشخیص موجودیت‌های نامدار، پیکره موجودیت‌های نامدار، پردازش زبان طبیعی، یادگیری ماشین، برچسب اجزای کلام.

۱- مقدمه

تشخیص موجودیت‌های نامدار[۱] (NER) یکی از مسائل پردازش زبان طبیعی[۲] است که کاربردهای عمده آن در سیستم‌های خلاصه‌ساز متون، استخراج اطلاعات، بازیابی اطلاعات، پرسش و پاسخ، مترجم ماشینی و دسته‌بندی اسناد است [۸]. وظیفه یک سامانه تشخیص موجودیت نامدار، علاوه بر تعیین مرز هر یک از موجودیت‌های نامدار، تشخیص نوع آن موجودیت و قرار دادن آن در دسته‌های از پیش تعیین شده است. این دسته‌ها شامل اسامی خاص افراد، مکان‌ها (شهر، کشور و غیره)، اسامی سازمان‌ها، عبارت‌های زمانی، کمیت‌ها، عبارت‌های پولی، درصد و غیره می‌شوند. به طور کلی تشخیص موجودیت نامدار با استفاده از روش‌های مبتنی بر قانون و یا روش‌های یادگیری ماشین انجام می‌شود. برای تولید سامانه‌های تشخیص موجودیت نامدار با روش یادگیری ماشین نیاز به یک پیکره استاندارد و مناسب وجود دارد. اگرچه چنین پیکره‌ای در زبان‌های مختلف از جمله انگلیسی تهیه شده است، اما در زبان فارسی تعداد چنین پیکره‌هایی انگشت‌شمار و غالباً حجم آنها محدود است.

این مقاله به بررسی روند تولید یک سامانه تشخیص موجودیت‌های نامدار مبتنی بر روش‌های یادگیری ماشینی برای زبان فارسی می‌پردازد. در این سامانه از برچسب اجزای کلام (POS) به عنوان یک ویژگی مفید در روش یادگیری ماشین استفاده شده است. این مقاله به بررسی اثر دقت این ویژگی بر دقت کل سامانه NER می‌پردازد. با توجه به اینکه ماژول برچسب‌گذار اجزای کلام نیز مبتنی بر یادگیری ماشین و دارای خطا است، نتایج به دست آمده از این مقاله نشان خواهد داد که در کاربردهای واقعی دقت سامانه NER تا چه حد به دقت ماژول برچسب‌گذار اجزای کلام وابسته است و به عبارت دیگر برچسب‌گذار اجزای کلام مورد استفاده در سامانه NER تا چه حد بایستی دقیق باشد.

در بخش دوم مقاله، ابتدا مروری بر کارهای پیشین صورت گرفته در تشخیص موجودیت‌های نامدار صورت می‌گیرد. در بخش سوم به معرفی پیکره مورد استفاده (پیکره اعلام) پرداخته می‌شود. بخش چهارم اختصاص به معرفی سامانه تشخیص موجودیت‌های نامدار دارد. این سامانه با استفاده از پیکره اعلام تهیه شده است. همچنین در این بخش به بیان نتایج حاصله از

[۱] Named Entity Recognition
[۲] Natural Language Processing

ارزیابی سامانه پرداخته می‌شود و در نهایت بخش پنجم به نتیجه‌گیری و ارائه پیشنهاداتی برای کارهای آتی تخصیص یافته است.

۲- مروری بر کارهای پیشین

تحقیق بر روی مساله تشخیص موجودیت نامدار به طور جدی در سال ۱۹۹۶ میلادی در رقابت MUC-6 مطرح شد. در این رقابت وظیفه شرکت‌کنندگان تشخیص و دسته‌بندی سه نوع اصلی موجودیت (ENAMEX، TIMEX و NUMEX) بود [۱۰]. پس از رقابت MUC-6 رخدادهای علمی زیادی با موضوع تشخیص موجودیت نامدار شکل گرفت. از جمله این رقابت‌ها می‌توان به MUC-7، IREX، CoNLL2002، CoNLL2003 اشاره کرد. دستاورد این رقابت‌ها علاوه بر به چالش کشیدن روش‌های تشخیص موجودیت نامدار، تولید پیکره‌های برچسب‌گذاری شده بود. علاوه بر پیکره‌های تولید شده در خلال این رقابت‌ها، پیکره‌های دیگری نیز در این زمینه تولید شده‌اند. این پیکره‌ها در تعداد و نوع برچسب‌ها با یکدیگر تفاوت دارند. بیشتر پیکره‌های تولید شده در این زمینه، طبق تعریف ارائه شده در رقابت CoNLL2003 [۱۵] دارای چهار موجودیت شخص، مکان، سازمان و متفرقه می‌باشند. تعداد پیکره‌های تولید شده در زبان فارسی بسیار محدود است. اخیراً پیکره‌ای برای زبان فارسی تولید و گزارش شده است که شامل ۴۰۰ هزار توکن از پیکره متنی فارسی [۵] است که به صورت دستی برچسب‌گذاری شده است. این پیکره تنها شامل سه برچسب شخص، مکان و سازمان است و برچسب‌های دیگر را در بر نمی‌گیرد. تعداد کل موجودیت‌های این پیکره در حدود ۱۵ هزار موجودیت است [۴]. همچنین شرکت تجاری appen[۱] در سال ۲۰۱۵ میلادی برای چند زبان از جمله زبان فارسی، عربی و انگلیسی پیکره‌های موجودیت‌های نامدار جداگانه با حدود ۵۰۰ هزار کلمه برای هر زبان منتشر کرده است. این پیکره‌ها تنها حاوی ۸ برچسب موجودیت شامل موجودیت‌های شخص، سازمان، مکان، ملیت، دین، عنوان، موجودیت‌های مکان‌های سیاسی و امکانات می باشند. همچنین پیکره اعلام با بیش از ۲۵۰ هزار توکن و ۱۳ برچسب موجودیت نامدار یک پیکره مناسب، جامع و استاندارد برای پردازش زبان طبیعی فارسی است [۲].

تشخیص موجودیت‌های نامدار، با روش‌های متفاوتی انجام می‌گیرد. از جمله این روش‌ها می‌توان به روش‌های استخراج با واژگان، روش‌های مبتنی بر یادگیری باسرپرست و یا نیمه سرپرستی شده و یا بدون سرپرستی اشاره نمود. علی‌رغم سهولت پیاده‌سازی، روش‌هایی که تنها از دیکشنری در تعیین موجودیت‌های نامدار استفاده می‌کنند، به دلیل در نظر نگرفتن محتوای متن[۲]، دقت شناسایی و طبقه‌بندی بالایی ندارند و معمولا به عنوان بلوک جانبی –و یا درون بلوک روش‌های مبتنی بر یادگیری– قرار می‌گیرند و به تنهایی استفاده نمی‌شوند. در برخی از روش‌ها از قوانین برای تعیین موجودیت‌های نامدار استفاده می‌شود [۹]. استفاده از روش‌های بدون سرپرست نیازی به پیکره متنی برچسب خورده جهت تعلیم ندارد، با این وجود برای ارزیابی آنها وجود داده متنی با برچسب موجودیت نامدار ضروری است.

بهترین عملکرد بین سامانه‌های تشخیص موجودیت اسمی به روش‌های باسرپرست مربوط می‌شود. در این روش‌ها یادگیری روی داده برچسب‌خورده انجام می‌شود. نمونه‌هایی از الگوریتم‌های به کاررفته در یادگیری با سرپرست عبارتند از: مدل مخفی مارکوف[۳] [۶]، روش مبتنی بر ماکزیمم آنتروپی[۴][۷]، درخت تصمیم[۵] [۱۴] و روش میدان‌های تصادفی شرطی[۶] [۱۱]. دلیل موفقیت این روش‌ها، توجه به محتوای متن است که در تشخیص عبارات مبهم، بهتر از سایر روش‌ها عمل می‌کنند.

در زبان فارسی نیز چند پژوهش در زمینه تولید سامانه برچسب‌گذار موجودیت‌های نامدار انجام گرفته است. با توجه به در دسترس نبودن داده مناسب و کافی در زبان فارسی، تمرکز اغلب این پژوهش‌ها بر روش‌های مبتنی بر قانون است [۳]. سادات‌مرتضوی در پژوهشی در زبان فارسی سامانه‌ای برای تشخیص موجودیت‌های نامدار و دسته‌بندی آنها در زبان فارسی تهیه کرده است. این سامانه با بکارگیری ساختار واژگانی اسامی خاص و نیز الگوهای متنی ممکن برای اسم‌های خاص متعلق به یک دسته، سعی در شناسایی موجودیت‌های نامدار می‌کند [۳].

در پژوهش دیگری [۱] با استفاده از چهار طبقه‌بندی کننده خطی، بیزین، نزدیکترین همسایگی[۷] و شبکه عصبی، سامانه تشخیص موجودیت‌های نامدار آموزش داده شده است. نتیجه این سامانه با استفاده از طبقه‌بندی کننده خطی و نزدیکترین همسایگی مقدار f-measure ۹۱٪ گزارش شده است.

در پژوهش دیگری در زمینه تولید سامانه تشخیص موجودیت نامدار برای زبان فارسی با استفاده از ترکیب روش یادگیری ماشین مدل مخفی مارکف و روش مبتنی بر قواعد سامانه‌ای تولید شده است که قادر به تشخیص سه موجودیت اسامی اشخاص، مکان و سازمان است. این سامانه بر روی حجم داده ۳۲ هزار کلمه‌ای تست شده است و نتیجه آن ۸۵,۹۳ درصد برای معیار f-measure گزارش شده است [۱۳].

۳- پیکره اعلام

پیکره مورد استفاده در سامانه تشخیص موجودیت‌های نامدار، «پیکره اعلام» است. این پیکره در پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی تهیه شده است. این پیکره بیش از ۲۵۰ هزار توکن دارد و تعداد برچسب‌های موجودیت نامدار به کار رفته در پیکره ۱۳ برچسب است.

۳-۱- دادگان خام مورد استفاده در پیکره اعلام

در تولید پیکره اعلام، از داده‌های موجود در بخش برچسب‌گذاری شده پیکره متنی فارسی (پیکره هشت میلیون کلمه‌ای) استفاده شده است تا از اطلاعات POS دقیق کلمات نیز استفاده شود. تعداد ۸۴۰۰ جمله به‌صورت تصادفی از این دادگان انتخاب شده است. انتخاب تصادفی موجب این می‌شود که مدل تعلیم‌یافته روی پیکره‌ای با این ویژگی، از قدرت تعمیم‌پذیری بالاتری برخوردار باشد. این تعداد جمله شامل بیش از ۲۵۰ هزار کلمه هستند. سپس جملات انتخاب‌شده یکبار دیگر مورد بازبینی قرار گرفته و یکسان‌سازی‌هایی در آنها صورت گرفته است [۲].

پیکره متنی فارسی[۸] یک پیکره استاندارد در زبان فارسی است. این پیکره شامل بیش از یکصد میلیون کلمه فارسی است که توسط پژوهشکده پردازش هوشمند علائم (این پژوهشکده از سال ۱۳۹۲ با مجوز رسمی وزارت علوم، تحقیقات و فناوری به پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی ارتقا یافته است) تهیه و جمع‌آوری شده است. تمام این داده‌ها که از منابع متنوع نوشتاری و گفتاری انتخاب شده‌اند، توکن‌بندی شده و توکن‌ها به‌صورت دستی تصحیح شده‌اند. بخشی از این پیکره که شامل هشت میلیون کلمه است علاوه بر توکن‌بندی و تصحیح دارای برچسب اجزای کلام (POS) در سطح کلمه است [۵].

۳-۲- برچسب‌های موجودیت نامدار در پیکره اعلام

در اکثر پیکره‌های موجودیت نامدار تولیدشده، تعداد برچسب‌ها مطابق دستورالعمل ارائه شده در رقابت CoNLL2003 شامل چهار برچسب اسامی شخص، اسامی مکان، اسامی سازمان‌ها و متفرقه می‌باشد [۱۵]. اما در پیکره اعلام تلاش بر تولید یک پیکره جامع بوده است که بتوان از سامانه تولید شده بر مبنای آن در اکثر وظایف پردازش زبان طبیعی فارسی استفاده نمود. بنابراین برای تحقق این هدف تعداد برچسب‌های تعریف شده در این پیکره افزایش یافته است و به ۱۳ برچسب رسیده است. افزایش تعداد برچسب‌ها، اطلاعات موجود در پیکره را افزایش می‌دهد و کمک می‌کند تا پیکره قابل تعمیم به مسائل مختلف پردازش زبان طبیعی باشد. در جدول ۱ نام این برچسب‌ها و تعریف مختصری از هر برچسب ارائه شده است.

برچسب‌گذاری جملات این مجموعه به‌صورت دستی انجام شده است. پس از تهیه دادگان و برچسب‌گذاری دستی آنها، دادگان به قالب استانداردِ IOB تبدیل شدند. در این قالب به هر کلمه موجود در متن یکی از سه برچسب B[۹]، I[۱۰] و یا O[۱۱] اختصاص داده می‌شود. کلمه آغازگر یک موجودیت، برچسب B، سایر کلمات موجودیت نامدار برچسب I و در نهایت باقی‌مانده کلمات برچسب O دارند. تعداد ۱۳ برچسب موجودیت نامدار به‌کار رفته در پیکره پس از تبدیل به قالب IOB به ۲۷ برچسب رسیده است [۲]. در جدول ۲ برچسب‌های موجودیت نامدار پیکره اعلام به همراه شمارش آنها در پیکره نمایش داده شده است.

---

۱ www.appen.com
۲ context
۳ Hidden markov model (HMM)
۴ Maximum Entropy (ME)
۵ Decesion tree (DT)
۶ Conditional random field (CRF)
۷ KNN
۸ Farsi Text Corpus
۹ Begin
۱۰ Inside
۱۱ Outside

۴- سامانه تشخیص موجودیت نامدار

جدول (۱): برچسب‌های موجود در پیکره اعلام

ردیف	موجودیت	تعریف
۱	شخص	نام و نام خانوادگی اشخاص مثال: «رستم دستان»، «عبید زاکانی»
۲	LNORP	مخفف چهار کلمه ملیت، دین، گروه سیاسی و زبان مثال:«ایرانی»، «جمهوری‌خواه»
۳	سازمان	شرکت‌ها، ادارات دولتی، نشریه‌ها و ... مثال:«روزنامه اطلاعات»، «اداره آموزش و پرورش»
۴	مکان	نام مکان‌های سیاسی و جغرافیایی مثال: «زعفرانیه»،«تبریز»، «کارون»
۵	رخداد	رخدادهای مهم تاریخی: مثال: «انقلاب اسلامی ایران»، «جام جهانی فوتبال»
۶	تاریخ	بیان کلی و جزئی از تاریخ: مثال:« ۲۲ بهمن ۱۳۵۷»
۷	بازه	زمان سپری شده یا بازه زمانی مثال: «ده سال گذشته»
۸	زمان	زمان‌های کوتاهتر از یک روز مثال: «ساعت ۶ صبح»
۹	درصد	عبارات شامل ٪ مثال: «۱۷٪»، «سی و هشت درصد»
۱۰	پول	عبارات پولی مثال: «۲۰۰ تومان»
۱۱	اندازه	اندازه‌گیری‌های استاندارد مانند: سن، مساحت، فاصله مثال: «۴۷ متر»، «صفر درجه سانتیگراد»
۱۲	عدد اصلی	شمارش عددی یا کمیت برخی از اشیاء: مثال:«سه»، «۷/۱»
۱۳	عدد ترتیبی	اعداد ترتیبی مثال: «نخستین»، «دوم»

جدول (۲): برچسب‌های موجودیت نامدار و شمارش هریک در پیکره اعلام.

ردیف	برچسب موجودیت	تعداد رخداد	درصد نسبت تعداد موجودیت به کل موجودیت‌ها
۱	سازمان	۲۶۳۲	۱۴,۱
۲	مکان	۴۹۰۴	۲۶,۳
۳	شخص	۲۸۶۳	۱۵,۴
۴	اصلی	۳۰۴۶	۱۶,۳
۵	تاریخ	۱۳۲۸	۷,۱
۶	LNORP	۱۴۱۳	۷,۶
۷	رخداد	۴۳۱	۲,۳
۸	بازه	۳۴۵	۱,۹
۹	اندازه	۳۸۸	۲,۱
۱۰	پول	۲۳۴	۱,۲
۱۱	ترتیبی	۶۷۷	۳,۶
۱۲	درصد	۲۰۰	۱,۱
۱۳	زمان	۱۸۳	۱,۰
۱۴	مجموع	۱۸۶۶۴	۱۰۰,۰۰

از پیکره اعلام در ساخت یک سامانه تشخیص موجودیت نامدار فارسی استفاده شده است. برای تعلیم مدل از روش‌های مبتنی بر یادگیری استفاده شده است. بهترین عملکرد بین روش‌های یادگیری ماشینی در این حوزه به روش CRF اختصاص دارد [۱۱]. این الگوریتم برپایه احتمال شرطی و نظریه گراف بنا شده است. برخلاف بسیاری از روش‌ها، مدل تصادفی شرطی تنها به یک نمونه بسنده نمی‌کند و برای برچسب‌دهی، بافتی که نمونه در آن قرار دارد را مورد توجه قرار می‌دهد. در تولید این سامانه از روش CRF و ابزار متن‌باز[۱] PocketCRF[۲] با زبان برنامه‌نویسی C++ استفاده شده است. برای آموزش مدل تشخیص موجودیت نامدار از دو ویژگی اصلی کلمه و برچسب اجزای کلام آن استفاده شده است. این ویژگی‌ها به صورت توابع ویژگی در مرحله آموزش به ابزار PocketCRF داده می‌شوند.

۴-۱ برچسب‌گذارهای اجزای کلام

برای برچسب‌گذاری اجزای کلام (POS)، از چهار ابزار مختلف برچسب‌گذار استفاده شده است. تمامی این برچسب‌گذارها با استفاده از بخشی از پیکره متنی که دارای برچسب POS است (پیکره ۸ میلیون کلمه‌ای) تعلیم داده شده‌اند و تفاوت آنها در تعداد و انواع برچسب‌های به کار رفته است.

برای تهیه ابزارهای برچسب‌گذاری POS مجموعه ۸ میلیونی پیکره متنی فارسی به دو قسمت آموزش و آزمون تقسیم شده است. قسمت آزمون شامل ۲۵۰ هزار کلمه است و مابقی داده‌ها به بخش آموزش اختصاص دارد. ابزارها با روش مبتنی بر مدل مخفی مارکوف تعلیم یافته‌اند [۱۲]. هر یک از این چهار ابزار دارای مجموعه برچسب متفاوتی هستند. این مجموعه برچسب‌ها به صورت زیر مجموعه‌هایی از برچسب‌های سلسله مراتبی موجود در پیکره متنی تهیه شده‌اند.

در ابزار برچسب‌گذار پایه از ۱۵ برچسب استفاده شده است. این برچسب‌ها شامل برچسب‌های مرحله اول پیکره متنی هستند. یعنی در واقع این ابزار مشخص‌کننده نوع کلی مقوله کلمه است. برای مثال اسم، فعل، صفت و ... .

در برچسب‌گذار کسره اضافه علاوه بر ۱۵ برچسب اصلی مقوله کلمات، برچسب کسره اضافه نیز استفاده شده است. لذا این ابزار علاوه بر تشخیص مقوله اصلی کلمات قادر به تشخیص کسره اضافه نیز هست. برای مثال: اسم، اسم+کسره اضافه، فعل، صفت، صفت+کسره اضافه و ... .

در ابزار برچسب‌گذار ۳۳ علاوه بر مقوله اصلی کلمه، جزئیاتی از قبیل جمع یا مفرد بودن اسم، نوع صفت، نوع قید، زمان فعل و ... نیز مشخص شده است. برای مثال اسم جمع، اسم مفرد، صفت ساده، فعل ماضی و ... .

در ابزار برچسب‌گذار ۱۰۰ از ویژگی‌های بیشتری در مجموعه برچسب استفاده شده است. علاوه بر ویژگی‌های موجود در برچسب‌گذار ۳۳، ویژگی‌های خاص و عام بودن اسم، وجود واژه‌بست در کلمه، زمان فعل، شخص فعل و ... نیز مشخص شده است.

جدول ۳ بیانگر دقت هر برچسب‌گذار به همراه تعداد برچسب‌های آنهاست. با توجه به نتایج موجود در جدول ۳ مشاهده می‌شود که ابزارهای برچسب‌گذاری POS دارای دقت‌های مناسبی هستند و برچسب‌گذار پایه با ۱۵ برچسب اجزای کلام از سایر برچسب‌گذارها دقت بیشتری دارد. لذا با استفاده از این برچسب‌گذار داده‌های پیکره اعلام برچسب‌گذاری می‌شوند و از این برچسب‌ها به عنوان ویژگی در CRF استفاده می‌شود. لازم به ذکر است که داده‌های مورد استفاده در بخش آموزش برچسب‌گذارها با داده‌های بخش آموزش پیکره اعلام همپوشانی ندارند و از اینرو مدل نهایی NER قابلیت تعمیم پذیری بیشتری خواهد داشت.

جدول (۳): خصوصیات ابزارهای برچسب‌گذاری اجزای کلام

برچسب‌گذار	تعداد برچسب‌ها	صحت[۳]
برچسب‌گذار پایه	۱۵	۹۸/۴
برچسب‌گذار کسره اضافه	۲۵	۹۷/۹
برچسب‌گذار ۳۳	۳۳	۹۷/۲
برچسب‌گذار ۱۰۰	۱۰۰	۹۴/۸

[۱] Open source

[۲] Sourceforge.net/projects/pocket-crf-1

[۳] accuracy

۴-۲ آموزش CRF با استفاده از ویژگی برچسب اجزای کلام

پیکره اعلام به دو بخش آموزش و آزمون تقسیم شده است. قسمت آموزش شامل ۹۰ درصد از پیکره است و ۱۰ درصد بقیه به بخش آزمون اختصاص دارد. این پیکره شامل کلمات، برچسب اجزای کلام و برچسب موجودیت نامدار است.

در آزمایش اول بدون استفاده از برچسب اجزای کلام، تنها با استفاده از ویژگی کلمات، CRF آموزش داده شده است و مدل تولید شده بر روی قسمت آزمون پیکره، آزمایش شده است.

در آزمایش دوم داده‌های پیکره اعلام با استفاده از بهترین برچسب‌گذار اجزای کلام معرفی شده در بخش قبل (برچسب‌گذار پایه)، برچسب‌گذاری شده‌اند. سپس یادگیرنده CRF با استفاده از ویژگی‌های کلمات و برچسب اجزای کلام بر روی این داده‌ها آموزش داده شده است و مدل تولید شده بر روی قسمت آزمون آزمایش شده است.

در آزمایش سوم از برچسب‌های POS دقیق موجود در پیکره متنی استفاده شده است به دلیل اینکه داده‌های پیکره اعلام از پیکره متنی برداشته شده‌اند دسترسی به برچسب دقیق اجزای کلام آنها میسر بود. لذا در این آزمایش CRF با استفاده از ویژگی برچسب دقیق POS آموزش داده شده است و مدل تولیدشده بر روی قسمت آزمون، آزمایش شده است.

نتایج هر سه آزمایش در جدول ۴ مشاهده می‌شود. با توجه به نتایج جدول ۴، مشاهده می‌شود که عدم استفاده از برچسب اجزای کلام منجر به پایین‌ترین دقت[۱] و فراخوانی[۲] شده است. بنابراین ویژگی POS کلمات در افزایش کارایی سامانه موثر است. بیشترین دقت در حالت استفاده از برچسب دقیق بوده است که مقدار f-measure ۸۵/۰ را کسب نموده است. از طرف دیگر عدم استفاده از برچسب دقیق POS (بوسیله ماژول برچسب‌گذار خودکار POS با صحت ۹۸/۴ درصد) نیز نتیجه شناسایی NER مناسبی در حد استفاده از ویژگی های دقیق POS کلمات نشان می دهد. بنابراین دقیق بودن و یا داشتن خطای اندک در تعیین برچسب های POS تاثیر قابل توجهی در دقت خروجی سامانه تشخیص موجودیت نامدار ندارد و به عبارت دیگر، روش یادگیری ماشین در سامانه تشخیص موجودیت نامدار قادر است خطاهای ایجاد شده توسط برچسب‌گذار اجزای کلام را فراگیرد.

جدول (۴): نتایج سامانه تشخیص موجودیت نامدار در حالت عدم استفاده از POS و با استفاده از POS دقیق و غیر دقیق

	دقت	فراخوانی	معیار F1
مدل NER بدون استفاده از POS	۸۳/۲	۶۰/۹	۷۰/۳
مدل NER با استفاده از POS دقیق	۹۲/۹	۷۸/۴	۸۵/۰
مدل NER با استفاده از POS غیر دقیق	۹۲/۳	۷۸/۶	۸۴/۹

۵- جمع‌بندی و نتیجه‌گیری

در این مقاله، نحوه تولید یک سامانه تشخیص موجودیت‌های نامدار بررسی شده است. این سامانه با استفاده از روش یادگیری ماشین میدان‌های تصادفی شرطی و پیکره اعلام –که یک پیکره استاندارد موجودیت‌های نامدار است–تهیه‌شده است. برچسب اجزای کلام یک ویژگی مورد استفاده در آموزش این سامانه بوده است. در این مقاله به تاثیر دقت این ویژگی در دقت نهایی سامانه پرداخته شده است. با بررسی نتایج این آزمایش نتیجه گرفته می‌شود که دقیق یا غیر دقیق بودن برچسب اجزای کلام به معنی استفاده از برچسب‌های دستی (موجود در پیکره متنی) و یا استفاده از ابزارهای برچسب‌گذاری تاثیر اندکی بر نتیجه نهایی سامانه تشخیص موجودیت‌های نامدار دارد و این به این معنی است که این سامانه قادر است خطاهای برچسب‌گذاری اجزای کلام را فراگیرد.

مراجع

[۱] اصفهانی، سیدعبدالحمید، راحتی قوچانی، سعید و جهانگیری، نادر، «**سیستم شناسایی و طبقه بندی اسامی در متون فارسی**»، پردازش علایم و داده‌ها، شماره ۱۳، صص. ۸۸-۷۷، ۱۳۸۹.

[۲] حسین‌نژاد، شادی، شکفته، یاسر، امامی آزادی، طاهره. «**پیکره اعلام: یک پیکره استاندارد موجودیت‌های نامدار برای زبان فارسی**»، پردازش علایم و داده‌ها، در دست داوری.

[۳] سادات‌مرتضوی، پونه، شمس‌فرد، مهرنوش. «**شناسایی موجودیت‌های نامدار در متون فارسی.**» پانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر، تهران، ۱۳۸۸.

[۴] عبدوس، محمد، مینایی بیدگلی، بهروز و قدمنان، حمیدرضا، «**تولید پیکره واحدهای اسمی فارسی.**» اولین همایش ملی زبان‌شناسی پیکره‌ای، تهران، ۱۳۹۴.

[5] Bijankhan, M., Sheykhzadegan, J., Bahrani, M. & Ghayoomi. M. **"Lessons from Building a Persian Written Corpus: Peykare."** *Language Resources and Evaluation*, 45(2), 143–164, 2011.

[6] Bikel, Daniel M.; Miller, S.,Schwartz, R., Weischedel, R. **"Nymble: a High-Performance Learning Name-finder".** in *In Proceedings of Conference on Applied Natural Language Processing*, 1997.

[7] Borthwick, A., Sterling, J.; Agichtein, E.; Grishman, R. **"NYU: Description of the MENE Named Entity System as used in MUC-7".** in In Proceedings of the Seventh Message Understanding Conference, 1998.

[8] Che, W., Wang,M. , Manning, C.D. and Liu, T. "**Named Entity Recognition with Bilingual Constraints."** In HLT-NAACL, pp. 52-62. 2013.

[9] Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F. and Vaithyanathan, S. **"Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks".** *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1002–1012, 2010.*

[10] Grishman, R, and Sondheim, B. **"Message Understanding Conference-6: A Brief History."** *COLING*. Vol. 96. 1996.

[11] McCallum, A., Li, W. **"Early results for named entity recognition with conditional random fields, feature Induction and Web-enhanced Lexicons".** in *Proceedings of CONLL, pages 188–191, 2003.*

[12] Halacsy, P., Kornai, A. and Oravecz, C. **"HunPOS: An Open Source Trigram Tagger."** Proceedings of the 45th annual meeting of the ACL on Interactive Poster and Demonstraition sessions. Association for Computational Linguistics, 2007.

[13] Moradi, H. and Ahmadi, F. **"A hybrud approach for Persian Named Entity Recognition".** 7th conference on information and knowledge Technology (IKT), Urmia, Iran, 2003.

[14] Sekine, S. and Nobata, C. **"Definition, Dictionaries and tagger for extended named entity Hierarchy".** in proceedings of conference on Language Resources and Evaluation., 2004.

[15] Tjong Kim Sang, E. F., and De Meulder, F. **"Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition."** Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003.

۱ precision

۲ recall