بهبود شناسایی موجودیت‌های نامدار فارسی با استفاده از کسرهٔ اضافه

محمد عبدوس[1]* و بهروز مینایی بیدگلی[2]

[1] دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران و آزمایشگاه پردازش و تحلیل متن شرکت آرمان رایان شریف، تهران، ایران

[2] دانشکده مهندسی کامپیوتر، دانشگاه علم وصنعت ایران، تهران، ایران

چکیده

تشخیص موجودیت‌های نامدار فرآیندی است که در آن اسامی اشخاص، مکان‌ها(شهرها، کشورها، دریاها و غیره)، سازمان‌ها(شرکت‌های خصوصی و دولتی، نهادهای بین‌المللی و غیره)، تاریخ، واحدهای پولی و درصدها در یک متن شناسایی می‌شوند. تشخیص موجودیت‌های نامدار نقشی اساسی در سامانه‌های پرسش و پاسخ، خلاصه‌سازی، ترجمه ماشینی، برچسب‌زن نقش معنایی، جستجوی معنایی، استخراج رابطه و شناسایی نقل قول دارند. در این مقاله ابتدا فرهنگ واژگان موجودیت‌های سازمان، مکان و اشخاص با استفاده از محتوای ویکی‌پدیای فارسی استخراج شد؛ سپس با استفاده از قواعد، سامانهٔ پیشنهادی توسعه یافت. در ادامه دقت شناسایی موجودیت‌های نامدار با استفاده از کسرهٔ اضافه که یکی از ویژگی‌های مهم زبان فارسی است، بهبود داده شد. جهت ارزیابی سامانه تعداد ۴۲ هزار کلمه از پیکرهٔ بی جن‌خان به‌صورت دستی برچسب زده شدند و معیار F ۸۱/۹۲ درصد به‌دست آمد. نتایج حاکی از آن است که با استفاده از کسرهٔ اضافه در سامانه‌های تشخیص موجودیت دقت آن‌ها به‌طور قابل ملاحظه‌ای افزایش می‌یابد.

واژگان کلیدی: تشخیص موجودیت‌های نامدار پردازش زبان طبیعی، مبتنی بر قاعده، ویکی‌پدیا، کسره اضافه

Improving Named Entity Recognition Using Izafe in Farsi

Mohammad Abdous*[1]& Behrooz Minaei Bidgoli[2]

[1] MSc in Artificial Intelligence at IUST. Tehran, Iran

[2] Iran University of Science and Technology

Abstract

Named entity recognition is a process in which the people's names, name of places (cities, countries, seas, etc.) and organizations (public and private companies, international institutions, etc.), date, currency and percentages in a text are identified. Named entity recognition plays an important role in many NLP tasks such as semantic role labeling, question answering, summarization, machine translation, semantic search, and relation extraction and quotation recognition systems. Named entity recognition in the Persian language is far more complex and more difficult than English. In English texts usually proper nouns begin with capital letters and this feature makes it easy to identify named entities, but this feature is absent in Persian language texts. To create a named entity recognition system, generally three methods are being used which include rule-based, machine-learning-based and hybrid methods. Each of these methods has its own advantages and disadvantages. Lack of named entity labeled data is the greatest challenge in Persian text. Because of this problem usually rule-based methods used to extract entities.

In this paper firstly, the dictionary of organizations, places and people were extracted from Wikipedia. Wikipedia is one of the best sources for extracting entities in which more than 200000 Farsi-named entities are known to exist. The proposed algorithm classify each Wikipedia article title by using its categories. Each

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

of Wikipedia titles has several categories that can be used to partially identify the named entity type. Then named entity recognition accuracy (precision) was increased using the rules. These rules can be divided into 3 categories that include morphological rules, adjacency and text patterns. The most important rules are adjacency rules. By using these rules the type of entity with the word nearby each entity (like Mr, Mrs , …) can be identified. To evaluate the system, 42000 tokens of BijanKhan corpus were manually annotated (labeled). Early F-measure was calculated 78.79 percent. In continue, named entity recognition accuracy (precision) improved using izāfe which is one of the important Persian language features and 81.94 percent for F-measure was achieved. The results showed that using izāfe in named entity recognition systems significantly increases their accuracy.

Keywords: Named Entity Recognition, Natural Language Processing, Rule Based, Wikipedia, Izafe

۱- مقدمه

دانش هوش مصنوعی نیازمند سامانه‌هایی است که بتوانند مانند انسان فکر و عمل کنند. در این دانش شاخه‌های مختلفی از علم گنجانده شده است که یکی از آن‌ها پردازش زبان طبیعی است. پردازش زبان طبیعی[1] شامل فرآیندهایی است که به رایانه قدرت تشخیص زبان طبیعی (به‌عنوان مثال زبان فارسی) را می‌دهد. اهمیت پردازش زبان طبیعی توسط رایانه، آنجایی نمود پیدا می‌کند که حجم عظیمی از اطلاعات غیرساخت‌یافته وجود دارد و برای استخراج اطلاعات از این متون نیاز به زمان و یا هزینه خیلی زیادی وجود داشته باشد؛ در اینجاست که دانش هوش مصنوعی به‌کمک انسان می‌آید و بسیاری از فرآیندهای پردازشی را به‌صورت خودکار و با زمان و هزینهٔ کم به سرانجام می‌رساند. در سال‌های اخیر در کشور ما نیز به‌خاطر حجم وسیع اطلاعات متنی که هرروزه انتشار می‌یابد، پردازش هوشمند مورد توجه قرار گرفته است. دانش پردازش زبان طبیعی دانشی است که بین دانش هوش مصنوعی و زبان‌شناسی قرار گرفته و نیاز است تا برای پردازش‌های زبان طبیعی از هر دو دانش استفاده شود.

جهت استخراج اطلاعات از هر متن نیاز است تا یک سری از پردازش‌ها بر روی آن‌ها اعمال شود. این پردازش‌ها را به دو قسمت پردازش‌های پایه و پیشرفته می‌توان تقسیم‌بندی کرد. وجود ابزارهای پایه، لازمهٔ تولید و استفاده از ابزارهای پیشرفته پردازش زبان است. ابزارهای پایه شامل تکواژساز[2]، یکسان‌ساز[3] و ریشه‌یاب[4] و ابزارهای پیشرفته شامل تشخیص مرجع ضمایر[5]، تشخیص موجودیت‌های نامدار[6]، استخراج رابطه[7] و امثال آن می‌توانند باشند. تشخیص موجودیت‌های نامدار بدین معناست که اسامی خاص در یک متن را بتوان تشخیص داد و آن‌ها را به رده‌های مشخصی دسته‌بندی کرد. این رده‌ها مکان، اشخاص، سازمان، مقادیر و کمیت‌های عددی، واحدهای پولی و غیره می‌توانند باشند که به‌طورمعمول شناسایی سه ردهٔ شخص، مکان و سازمان به‌دلیل چالش‌های شناسایی آن‌ها از اهمیت ویژه‌ای برخوردار است. برای استخراج این اسامی از متن، روش‌های مختلفی از جمله روش‌های باقاعده، مبتنی بر یادگیری ماشین و روش‌های ترکیبی وجود دارد [10]. در روش‌های مبتنی بر قاعده با کشف الگوها یا بافت‌های زبانی و اعمال قواعد، موجودیت‌های نامدار شناسایی می‌شوند. به‌عنوان مثال بعد از شاخص‌هایی که شغل اشخاص را تعیین می‌کنند، مثل «دکتر، مهندس، استاد و نظیر آن» به‌احتمال زیاد یک موجودیت نامدار قرار می‌گیرد؛ یا در ادامه جمله «شرکت‌هایی مانند ...، ...، ... و ....» نام شرکت قرار می‌گیرد. با شناسایی این الگوها موجودیت‌های نامدار شناسایی می‌شوند. مزیت این نوع روش‌ها سادگی، دقت به‌نسبه خوب و سرعت بالا در پیاده‌سازی است؛ ولی به‌دلیل ویژگی پراکندگی ترتیبی اجزا و گروه‌های نحوی در زبان فارسی و همچنین استفاده از اصطلاحات متنوع زبانی، پوشش خوبی ندارند. در برخی از روش‌های مبتنی بر قاعده از فرهنگ لغت یا واژگان استفاده می‌کنند. در این نوع روش‌ها فهرستی از موجودیت‌های نامدار عینی زبان تهیه می‌شود. این فهرست از منابع مختلف مانند ویکی‌پدیا، فرهنگ لغات مخصوص، فهرست اسامی شرکت‌های یک کشور و حتی دنیا و امثال آن

[1] Natural Language Processing
[2] Tokenizer
[3] Normalizer
[4] Lemmatizer
[5] Coreference Resolution
[6] Named Entity Recognition
[7] Relation Extraction

می‌تواند استخراج شود. تولید این فهرست نیازمند شناخت منابع و در عین حال خلاقیت و ابتکار در به‌دست‌آوردن آن‌هاست. از چالش‌های اصلی این روش این است که دامنهٔ پوشش فهرست واژگان به دامنهٔ موجودیت‌های نامدار موجود در فهرست وابسته است؛ و هرقدر فهرست کامل‌تری وجود داشته باشد، پوشش بیشتری نیز حاصل خواهد شد. در این نوع روش‌ها مسائلی هم‌چون نحوهٔ برچسب‌گذاری متون با استفاده از فهرست، نحوهٔ برخورد با کوتاه شدهٔ موجودیت‌های نامدار (به‌عنوان مثال در بسیاری از موارد ممکن است، به جای «وزارت رفاه، کار و امور اجتماعی» فقط گفته شود «وزارت رفاه») مطرح است. با طولانی‌شدن فهرست برای افزایش پوشش مسئله بهره‌وری نیز مطرح می‌شود. در این روش لازم است از برخی گزینه‌های موجودیت‌های نامدار رفع ابهام شود؛ به‌عنوان مثال واژه «روحانی» می‌تواند به یک شخص معین یا به کلمه عام روحانی اشاره کند و این رفع ابهام با توجه به محتوای متن قابل انجام است.

در روش‌های یادگیری ماشینی از قواعد و فرهنگ لغات استفاده نمی‌شود و به جای آن از حجم زیادی دادهٔ برچسب‌خورده یا بدون برچسب (بسته به روش انتخابی) برای این کار استفاده می‌شود. به‌طورمعمول در این روش‌ها نیاز است تا پیکره‌ای برای فرآیند یادگیری وجود داشته باشد تا بتوان عملیات یادگیری را با استفاده از آن انجام داد. چالش اصلی استفاده از این روش‌ها عدم وجود پیکره است که این چالش در زبان فارسی بیشتر نمایان است. با توجه به عدم وجود دادگان تشخیص موجودیت‌های نامدار فارسی در این مقاله با استفاده از روش‌های مبتنی بر قاعدهٔ سامانهٔ تشخیص موجودیت نامدار ایجاد شد. یافتن منبع واژگان موجودیت نامدار یکی از چالش‌های روش مبتنی بر قاعده است که با استفاده از ویکی‌پدیای فارسی این چالش نیز برطرف شد. در ادامه از کسرهٔ اضافه که یکی از مهم‌ترین ویژگی‌های زبان فارسی است، برای بهبود سامانهٔ تشخیص موجودیت نامدار مورد استفاده قرار گرفته است.

۲- کارهای مرتبط

بیش‌ترین کارهایی که در زمینه تشخیص موجودیت نامدار در جهان انجام شده، مربوط به زبان انگلیسی است و به‌علت وجود پیکرهٔ انگلیسی، الگوریتم‌های مبتنی بر یادگیری ماشین سهم عمده‌ای را در روش‌های تشخیص موجودیت نامدار ایفا می‌کنند. یکی از نخستین پژوهش‌ها که در زمینهٔ تشخیص موجودیت نامدار صورت پذیرفته مربوط به کار *لیزا راو*[1] است. پژوهش‌های وی در سال ۱۹۹۱ در هفتمین کنفرانس هوش مصنوعی IEEE[2] ارائه شده است. در این مقاله یک سامانه جهت استخراج و شناسایی اسامی و طبقه‌بندی آن‌ها با استفاده از توابع اکتشافی و تعدادی قواعد دستی معرفی شد [12]. پس از این مقاله تا سال ۱۹۹۵ چندین مقالهٔ دیگر در زمینهٔ تشخیص موجودیت نامدار ارائه شد. در بیش‌تر این مقالات از تعدادی فهرست مربوط به اسامی سازمان‌ها، افراد، مکان‌ها و نظیر آن در شناسایی و طبقه‌بندی اسامی استفاده می‌شود و در این سامانه‌ها بین دقت سامانه و غنای فهرست‌های مورد استفاده رابطه مستقیمی وجود دارد. در سال ۱۹۹۹ سامانه‌ای توسط *آندری میخو*[3] معرفی شد که در آن از فهرست اسامی استفاده نشده بود. وی با استفاده از قواعد و مدل‌های آموزش‌دیده و بدون استفاده از واژگان توانست به فراخوانی ۶۱ و دقت ۸۲ برسد [11]. *دیمترا*[4] از سامانه‌های تشخیص موجودیت نامدار در متون اقتصادی یونان استفاده کرده است. این سامانه اسامی را بر اساس قواعد زبانی، شناسایی و دسته‌بندی می‌کند[8]. *های‌لونگ*[5] در شناسایی و طبقه‌بندی اسامی سامانه‌ای را معرفی کرد که از ویژگی‌های عمومی در طبقه‌بندی استفاده می‌کرد [4]. در بین زبان‌هایی که به زبان فارسی مشابهت دارند نیز *بنجامین فاربر*[6] با همکاری *نزار حبشه*[7] از یک برچسب‌زن ریخت‌ساختی[8] برای بهبود سامانه‌های تشخیص موجودیت نامدار در زبان عربی استفاده کرد که باعث کاهش ۱۴٪ خطا در این سامانه‌ها شد [7]. *خالد شعلان*[9] هم یک سامانه‌ای را جهت شناسایی اسامی افراد در زبان عربی ارائه کرد که بر پایه قواعد زبان عربی عمل می‌کرد [13]. *علی الصبا*[10] برای شناسایی اسامی خاص در زبان عربی از یک

[1] Rau
[2] Institute of Electrical and Electronics Engineers
[3] Mikheev
[4] Dimetra
[5] Leong
[6] Farber
[7] Nizar Habash
[8] morphological
[9] Shaalan
[10] Elsebai

تابع اکتشافی استفاده کرده است که اسامی افراد را در متون عربی شناسایی و استخراج می‌کند [6].

در زبان فارسی نیز *اصفهانی* و همکارانش بر روی تشخیص اسامی در زبان فارسی تلاش کرده و برای استخراج ویژگی از تابع N-gram استفاده کرده‌اند و سپس چهار رده‌بند خطی، بیزین، نزدیک‌ترین همسایه و شبکه عصبی را آموزش داده و به این نتیجه رسیده‌اند که با استفاده از شبکه عصبی نتایج بسیار مناسبی در جداسازی اسامی مکان و افراد از بقیهٔ اسامی حاصل می‌شود (٪۹۹). رده‌بند KNN[1] و خطی به‌طور میانگین اسامی مکان و افراد و اسامی عمومی را با F-measure ٪۹۱ رده‌بندی کرده است. مجموعهٔ آموزشی که آن‌ها برای این کار در نظر گرفته‌اند، پیکرهٔ برچسب خورده *بیجن‌خان* است که در آن اسامی خاص و مکان‌ها با برچسب سطح دوم مشخص شده و الزاماً موجودیت نامدار نیستند. برای نمونه در پیکرهٔ بی‌جن‌خان کلماتی مانند کشور، دانشگاه‌ها، اینجا، آنجا، اداره، دادگاه و بسیاری موارد دیگر به‌عنوان مکان برچسب خورده‌اند که در تعریف موجودیت نامدار نمی‌گنجد. همچنین اسامی اشخاص به‌تنهایی برچسب موجودیت نگرفته‌اند و در مجموعه‌ی عظیمی از اسامی خاص گنجانده شده است. در مورد سازمان‌ها نیز به‌هیچ‌وجه برچسبی به این نام در این پیکره وجود ندارد. این نکته به این معنا است که از پیکرهٔ بی‌جن‌خان به‌عنوان پیکرهٔ یادگیری تشخیص موجودیت نامدار نمی‌توان استفاده کرد و با توجه به دلایل ذکرشده خروجی کار را به‌عنوان خروجی سامانه تشخیص موجودیت نامدار نمی‌توان قلمداد کرد [1].

همچنین *مرتضوی* و همکارانش به معرفی سامانه‌ای توسعه‌یافته به‌منظور تشخیص موجودیت‌های نامدار و دسته‌بندی آن‌ها در زبان فارسی پرداخته‌اند. این سامانه با به‌کارگیری ساختار واژه‌ای اسامی خاص و نیز الگوهای متنی برای اسم‌های خاص متعلق به یک دسته، سعی در شناسایی موجودیت‌های نامدار می‌کند؛ علاوه‌براین با به‌کارگیری برچسب نحوی و معنایی برای هر کلمه و توجه به سایر رخدادهای آن در متن، عملیات رفع ابهام را برای بهبود شناسایی انجام می‌دهد. آن‌ها سامانه را بر اساس داده‌های آزمون که به‌صورت دستی برچسب‌زنی شده بود، مورد ارزیابی قرار دادند و به درصد دقت ۷۲ و فراخوانی ۷۶ و معیار F ۷۳/۹۴ درصد دست یافته‌اند [2].

[1] k-nearest neighbors

۳- چالش‌های زبان فارسی

در بیش‌تر زبان‌های لاتین و غربی کلماتی که با حروف بزرگ نوشته می‌شوند، نامزدهای مناسبی جهت اسامی خاص هستند. در این زبان‌ها تفاوت در شکل حروف (بزرگ یا کوچک بودن) در شناسایی اسامی خاص به سامانه کمک می‌کند. بسیاری از سامانه‌های تشخیص موجودیت نامدار در زبان انگلیسی به همین روش، عملیات شناسایی را انجام می‌دهند. برای نمونه آزمایشگاه پردازش زبان طبیعی دانشگاه استنفورد یکی از برترین آزمایشگاه‌های جهان است که در زمینه پردازش زبان طبیعی از جمله بحث تشخیص موجودیت‌های نامدار فعالیت کرده است. این آزمایشگاه سامانهٔ تشخیص موجودیت نامدار را به زبان‌های انگلیسی، آلمانی و چینی توسعه داده است. همچنین نسخه برخطی[2] دارد که پس از دریافت متن ورودی و پردازش آن، موجودیت‌های نامدار شناسایی و استخراج می‌شوند. یکی از مشکلات اصلی این سامانه حساس‌بودن به حروف کوچک و بزرگ است؛ به‌عنوان مثال اگر موجودیت‌های نامدار با حرف کوچک به سامانه وارد شوند، سامانه قادر به شناسایی آن‌ها نخواهد بود[3]. مشکل دیگری که در شناسایی و طبقه‌بندی اسامی متون فارسی وجود دارد، مربوط به رسم‌الخط زبان فارسی است. این رسم‌الخط به تأیید فرهنگستان ادب و زبان فارسی رسیده و در آن اغلب پیشوندها و پسوندها به‌صورت جدا از هم نوشته می‌شوند که این امر باعث عدم شناسایی درست حد و مرز یک کلمه شده و پیشوندها و پسوندهای یک کلمه را به‌درستی نمی‌توان شناسایی کرد. این مشکل شناسایی ابتدا و انتهای موجودیت را با چالش روبه‌رو می‌کند. هرچند که فرهنگستان ادب و زبان فارسی نیم‌فاصله را برای حل این مسئله پیشنهاد کرده، اما با توجه به عدم رعایت آن در بسیاری از متون، هنوز این مشکل پابرجاست. موضوع دیگری که سامانه‌های تشخیص موجودیت نامدار فارسی با آن مواجه‌اند، عدم وجود مجموعه داده قوی و غنی برای آموزش است. سامانه‌های شناسایی و طبقه‌بندی موجودیت نامدار از روش‌های یادگیری ماشین و الگوریتم‌های با ناظر نیز استفاده می‌کنند و وجود یک مجموعه آموزش برای فرآیند یادگیری لازم به نظر می‌رسد. وجود چنین مجموعه داده‌ای به بالابردن صحت و دقت سامانه‌کمک شایانی

[2] http://nlp.stanford.edu:8080/ner/process

[3] iran برای نمونه

می‌تواند کند.

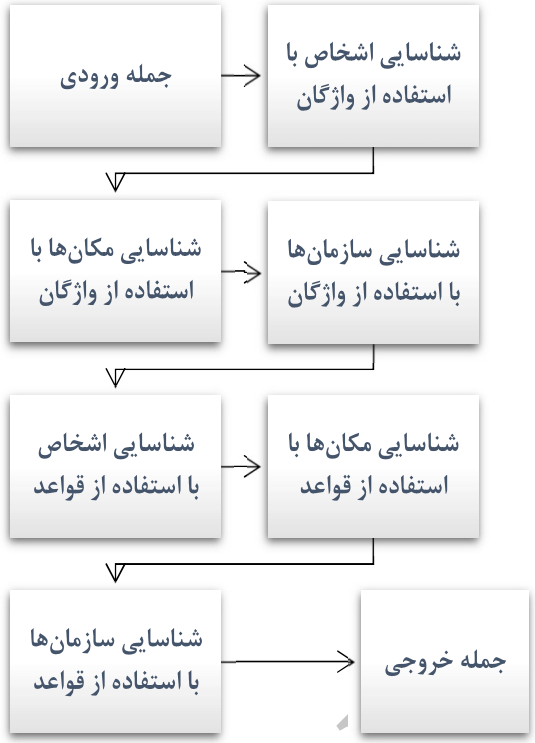
سامانه‌های تشخیص موجودیت نامدار وابستگی زیادی به متن و زبان دارند و ممکن است، نتوان آن را برای یک زبان نوشت و بدون اعمال تغییراتی به زبان‌های دیگر تعمیم داد. البته لازم به ذکر است در مواردی نظیر شبکه‌های عصبی عمیق قابلیت تعمیم روش وجود دارد. در ادامه پاره‌ای از چالش‌های زبان فارسی بیان می‌شود. برخی از مشکلات مطرح‌شده ویژگی‌های خاص زبان فارسی است:

- حذف نقش‌نمای اضافه در نوشتار فارسی که برخی از کاربردهای پردازش متن با آن مواجه‌اند. البته تأثیرگذارترین واکه‌ای که در متون فارسی نمایش داده نمی‌شود، کسره اضافه است. حذف کسرهٔ اضافه در نوشتار منجر به ایجاد مشکل در تشخیص مرزهای عبارات اسمی می‌شود.
- هرگز تمامی اسامی خاص یک زبان را نمی‌توان مشخص کرد و آن‌ها را در فهرست‌های مشخصی قرار داد؛
- بسیاری از اسامی اشخاص معنای صفتی نیز دارند (مانند پارسا). این مشکل در روش‌های مبتنی بر قواعد نمود بیشتری پیدا می‌کند؛
- موجودیت‌های نامدار در متون فارسی برخلاف زبان انگلیسی با نشانهٔ خاصی برجسته نمی‌شوند؛
- عدم وجود پیکرهٔ متنی تشخیص موجودیت مناسب در زبان فارسی؛
- چندمعنایی بعضی از کلمات در زبان فارسی مثل کلمه« شیر». برای شناسایی موجودیت‌ها نیاز به تحلیل معنایی است؛
- بیش‌تر جملات زبان فارسی دارای ساختار فاعل-مفعول-فعل هستند؛ اما بیشتر اوقات این ترتیب رعایت نمی‌شود. برای مثال «من دیروز در مدرسه کتاب را به علی دادم» یا «دیروز کتاب را در مدرسه به علی دادم». در صورت استفاده از برچسب‌های وابستگی در تشخیص موجودیت نامدار این چالش نمایان می‌شود؛
- عدم تطابق عناصر جمله: «آقای مدیر آمدند» درصورتی‌که کلمه «آقای» شاخص موجودیت شخص است.

با توجه به مشکلات ذکرشده به ساده‌تربودن فرآیند تشخیص‌موجودیت‌های نامدار در زبان انگلیسی می‌توان پی‌برد.

۴- روش پیشنهادی

روش پیشنهادی برای تشخیص موجودیت‌های نامدار شیوهٔ مبتنی بر قواعد به همراه فهرست واژگان است. با توجه به لزوم وجود ترتیب در شناسایی موجودیت‌های نامدار باید اولویت شناسایی آن‌ها مشخص شود. در شکل (۱) این اولویت به نمایش درآمده که براساس آن ابتدا تشخیص با استفاده از فهرست لغات انجام و در ادامه روش مبتنی بر قاعده برای پوشش موجودیت‌های نامدار استخراج نشده یا اصلاح نوع و دامنهٔ موجودیت‌های نامدار شناسایی‌شده استفاده می‌شود. از برتری‌های روش پیشنهادی این است که از سرعت و دقت قابل قبولی برخوردار است.



(شکل-۱): اولویت شناسایی موجودیت‌های نامدار
(Figure-1): Named Entity Recognition Priority

در ادامه برچسب‌های موجودیت‌های نامدار مورد استفاده در سامانه توسعه داده شده، معرفی می‌شوند.

۴-۱- برچسب‌های موجودیت نامدار

همان‌طورکه در کنفرانس آموزش محاسباتی زبان طبیعی در سال ۲۰۰۳ گزارش شد، پیکرهٔ متنی برچسب‌گذاری‌شده باید شامل کلمات متن به همراه برچسب مربوط به آن کلمه باشد[14]. کلاس‌های سازمان، مکان و شخص به عنوان رده های اصلی در نظر گرفته و سایر مواردی که در این سه دسته قرار نمی‌گیرند نیز با O معرفی شده‌اند؛ بنابراین هر کلمه باید با یکی از برچسب‌های زیر تعیین شود:

B-PERS: شروع یک نام خاص از اشخاص

I-PERS: میانه یا ادامهٔ یک نام خاص از اشخاص
B-LOC: شروع نام یک مکان
I-LOC: میانه یا ادامهٔ نام یک مکان
B-ORG: شروع نام یک سازمان
I-ORG : میانه یا ادامهٔ نام یک سازمان
O: کلمه‌ای که یک موجودیت نامدار نباشد.

۵- استخراج فهرست واژگان

در این قسمت با توجه به تعریف موجودیت‌های نامدار در بخش مقدماتی فهرستی از موجودیت‌های نامدار عینی زبان که از منابع مختلف مانند ویکی‌پدیا، فرهنگ لغات مختلف، فهرست اسامی شرکت‌های یک کشور و حتی دنیا و امثال آن می‌تواند استخراج شود، معرفی می‌شود. دامنه پوشش در روش مبتنی بر واژگان به دامنه موجودیت‌های نامدار موجود در فهرست وابسته بوده و هرقدر فهرست کامل‌تری وجود داشته باشد، پوشش بیشتری نیز حاصل خواهد شد. در این نوع روش‌ها موضوعاتی مانند نحوه برچسب‌گذاری متون با استفاده از فهرست و یا نحوه برخورد با کوتاه‌شدهٔ موجودیت‌های نامدار (به‌عنوان مثال در بسیاری از موارد ممکن است به جای «وزارت رفاه، کار و امور اجتماعی» صرفاً گفته شود «وزارت رفاه») مطرح است. به‌عنوان مثال واژه «روحانی» می‌تواند به یک یا چند شخص معین یا به کلمه عام روحانی اشاره کند و لازم است این رفع ابهام انجام شود. بسیاری از این ابهام‌ها را با استفاده از قواعد می‌توان برطرف کرد که در بخش ۶ به آن اشاره شده است.

۵-۱- استفاده از ویکی‌پدیا به‌عنوان منبع واژگان

ویکی‌پدیا دانش‌نامه‌ای همگانی و آزاد است؛ بدین معنی که همگان قادر به نوشتن و ویرایش نوشتارهای موجود در آن هستند. البته این نوشتارها و ویرایش‌ها باید مطابق با اساس‌نامه ویکی‌پدیا باشند؛ یعنی مطالب بی‌طرفانه و بدون پایمال‌کردن حق نشر دیگران نوشته شده باشند. در ویکی‌پدیای فارسی تا سال ۲۰۱۶ تعداد ۴۶۲۲۷۶ مقاله وجود دارد و هر روز هزاران کاربر از این دانش‌نامه بازدید می‌کنند[1]. در زبان‌هایی مانند فارسی، پیکرهٔ تشخیص موجودیت استاندارد و قابل دسترس مهیا نشده و همین موضوع باعث روی‌آوردن پژوهش‌گران به دانش‌نامه‌های آزاد شده است. یکی از بهترین منابع واژگانی جهت استخراج موجودیت‌های نامدار، دانشنامه ویکی‌پدیا است [5]. هر مقالهٔ ویکی‌پدیا دارای عنوان و رده‌های مختلفی است. بسیاری از عناوین این صفحات جزو سه دسته مکان، سازمان و اشخاص نمی‌شوند و با ایجاد الگوریتمی باید این دسته‌ها را شناسایی کرد. مشکل اصلی در استفاده از ویکی‌پدیا وجود اسامی مشهور در آن است و اگر اسمی در ویکی‌پدیا وجود نداشته نباشد، امکان شناسایی آن وجود ندارد؛ یعنی شناسایی اسامی خاص، به شناسایی اسامی خاص مشهور محدود می‌شود و اگر آن اسم مقاله‌ای را در ویکی‌پدیا به خود اختصاص داده باشد، به‌عنوان موجودیت نامدار شناسایی می‌شود؛ به همین دلیل در کنار استفاده از فهرست واژگان از قواعد نیز کمک گرفته شده است.

۵-۲- استخراج موجودیت‌های نامدار از ویکی‌پدیا

جهت استخراج موجودیت‌های نامدار از ویکی‌پدیا نیاز به آشنایی با ساختار عنوان‌ها و رده‌های ویکی‌پدیا است. الگوریتم پیشنهادی برای دسته‌بندی عنوان‌های مقالات ویکی‌پدیا از رده‌های هر عنوان کمک می‌گیرد. هر یک از عنوان‌های ویکی‌پدیا به رده‌هایی نسبت داده شده‌اند که با استفاده از آن‌ها نوع موجودیت نامدار را تا حدی شناسایی و در منبع واژگانی می‌توان اضافه کرد. برای نمونه رده‌های فردی مانند علی اسماعیلی عبارت‌اند از: افراد زنده، زادگان ۱۹۷۷ (میلادی)، وزنه‌برداران قدرتی اهل ایران که با استفاده از این رده‌ها می‌توان اسامی خاص را در ویکی‌پدیا شناسایی کرد. برای ایجاد فهرست واژگان نیاز است تا داده‌های ویکی‌پدیای فارسی را به‌دست آورد. در بازه‌های زمانی مختلف وبگاه ویکی‌پدیا نسخهٔ پشتیبانی از ویکی‌پدیای فارسی را در دسترس عموم قرار می‌دهد. بعد از بارگیری آن‌ها نیاز به پیش‌پردازش‌هایی برای استفاده از این اطلاعات است؛ زیرا فرمت داده‌ها به‌صورت varbinary است و لازم به تغییر برای استفاده در زبان فارسی به varchar است. این داده‌ها، آماده بارگزاری به پایگاه داده هستند؛ سپس می‌توان با استفاده از پرس‌وجوهای[2] متنوع که شامل الگوی رده‌ها باشند، انواع موجودیت‌ها را استخراج کرد. با استفاده از

[1] https://fa.wikipedia.org/wiki/ویکی‌پدیا:آشنایی_با_دانشنامه

[2] Queries

رده‌های ویکی‌پدیا فهرست اشخاص به‌دست می‌آید. رده‌های معروفی که اسامی اشخاص در آن قرار داشتند، عبارت‌اند از: اهالی ....(مکان)، زادگان ....(مکان یا سال)، درگذشتگان ....(سال) که با استفاده از این رده‌ها مجموعه‌ای بیش از سی هزار اسم شخص یکتا به‌دست آمد. همین روند برای سازمان‌ها و مکان‌ها نیز انجام پذیرفت و برای مکان‌ها سه هزار اسم و برای سازمان‌ها دو هزار موجودیت نامدار استخراج شد؛ سپس جهت پالایش مجموعه، دوباره داده‌های واکشی‌شده فیلتر و بعضی از نویسه‌های عربی نیز یکسان‌سازی و اصلاح شد تا بتوان به‌عنوان فهرست واژگان در سیستم مطرح شود. برخی از رده‌های مورد استفاده در جدول (۱) نمایش داده شده است:

(جدول-۱): رده‌های مورد استفاده برای استخراج موجودیت‌های نامدار
(Tabel-1): Named Entity Categories

نوع	رده‌های مطرح	تعداد
شخص	اهالی .... زادگان ... درگذشتگان .... نویسندگان .... فیلسوفان.... افراد زنده	33000
مکان	شهرهای .... استآن‌هایی .... روستاهای ... ایستگاه‌های ...	3000
سازمان	سازمان‌های ... اداره‌های ... شرکت ... بنیان‌گذاری‌های .... شهرداری‌های ... فرمانداری‌های ...	2000

۶- شناسایی موجودیت‌های نامدار با استفاده از قواعد

در این روش با استفاده از قواعد، موجودیت‌های نامدار شناسایی می‌شوند. این قواعد را به سه دسته می‌توان تقسیم‌بندی کرد که عبارتند از قواعد ریخت‌شناسی، همجواری و الگوهای متنی. بسیاری از موجودیت‌های نامدار دارای برچسب اسم و صفت هستند و استفاده از برچسب اجزای سخن دقت شناسایی آن‌ها را می‌تواند بالا ببرد؛ لذا با استفاده از برچسب اجزای سخن نیاز به پردازش‌های کمتری برای استخراج موجودیت‌های نامدار بوده و دقت شناسایی آن‌ها نیز افزایش پیدا می‌کند.

۶-۱- قواعد ریخت‌شناسی

در این قواعد به شکل ظاهری کلمه توجه می‌شود و کلمات مجاور و یا برچسب آن‌ها در امر شناسایی موجودیت‌های نامدار دخیل نمی‌شوند. این دسته از قواعد عبارت‌اند از:

وجود اعداد یا برخی حروف انگلیسی یا نویسه‌های خاص مانند- ، . ،»،«،+ در کلمه که نشانه‌ای از موجودیت نامدار بودن آن کلمه است. مثال:

محور تهران-مشهد/ سفارت ج.ا.ا در مسکو/ سی.ان.ان /
هواپیمای اف ۱۶/شبکه BBC /گروه ۵+۱

با مشاهده برخی از وندها احتمال موجودیت نامداربودن کلمه افزایش پیدا می‌کند. بعضی از این پسوندها منجر به اسم خاص اشخاص و بعضی دیگر منجر به مکان می‌شوند. این پسوندها عبارت‌اند از:

پسوندهایی که باعث تولید اسم مکان می‌شوند: گاه، خانه، زار، ستان، آباد، دان، سار، سرا، شهر

مثال: ارمنستان، حسن‌آباد، نوشهر، قدمگاه، گرمسار، کوهسار

پسوندهایی که منجر به تولید اسم خاص افراد می‌شوند: پور، زاده، نژاد، مقدم، خان

۶-۲- قواعد همجواری

در این قواعد با توجه به کلمات مجاور یک اسم، اقدام به شناسایی موجودیت‌های نامدار می‌شود. همچنین نوع موجودیت را با استفاده از کلمه مجاور می‌توان شناسایی کرد. به این کلمات عنوان شاخص را می‌توان اطلاق کرد. در ادامه فهرستی از کلمات پرتکرار مجاور هر نوع موجودیت، گردآوری شده است:

موجودیت انسان (PER): سرهنگ، ژنرال، سرلشکر، دکتر، مهندس، مرحوم، دکتر، حاج، خانم، سردار، سرتیپ، پروفسور، شیخ، دریادار، امیر، حضرت، آقای، جناب، خانم، آیت‌الله، امام، سید، شهید، ماموستا

همچنین اگر بعضی از عبارات بعد از یک کلمه قرار گیرند آن کلمه موجودیت شخص می‌شود مثلاً: رضی الله عنه، صلی‌الله علیه، رحمه الله علیه، سلام الله علیها

موجودیت نامدار مکان (LOC): شهر، استان، ایالت، بزرگراه، چهارراه، میدان، پل، اتوبان، خیابان، کوچه، بن‌بست،

شهرک، منطقه، قاره، فرودگاه، ترمینال، دریاچه، صحرای، باغ، مسجد، امامزاده، مرقد، دانشگاه، کتابخانه، مدرسه، هتل، ساختمان، سد، جاده، رشته‌کوه‌های، رودخانه و امثال آن.

مثال: خیابان دکتر شریعتی

همچنین می‌توان تقسیم‌بندی دقیق‌تری را برای موجودیت نامدار مکان در نظر گرفت. برای مثال می‌توان مکان‌ها را به مکان‌های طبیعی، مذهبی، شهری و ... تقسیم‌بندی کرد.

می‌توان با استفاده از بعضی از سمت‌هایی که به اسم مکان ختم می‌شوند اسامی مکان را در متن شناسایی کرد. نمونه‌ای از این کلمات عبارت‌اند از: سفیر، شهردار، استاندار، فرماندار، بخشدار، دهیار

موجودیت سازمان (ORG): وزارت، سازمان، اداره، مرکز، بانک و امثال آن.

برای تشخیص موجودیت سازمان بعضی از سمت‌ها نیز در نظر گرفته‌شده‌اند و اگر اسمی بعد از این سمت‌ها بیاید به احتمال زیاد موجودیت سازمان است.

سمت‌های سازمان: مدیر، رئیس، وزیر، روابط عمومی، دبیرکل

جدول (۲) شاخص‌های انواع موجودیت نامدار برای قواعد همجواری نمایش داده شده است:

(جدول-۲): شاخص‌های همجواری
(Tabel-2): Named Entity Indicators

	شاخص	مثال
شخص	سرهنگ، ژنرال، سرلشکر، دکتر، مهندس، مرحوم، دکتر، حاج، خانم، سردار، سرتیپ، پروفسور، شیخ، دریادار، امیر، حضرت، آقای، جناب، خانم، آیت‌الله، امام، سید، شهید، ماموستا	امام خمینی، جناب آقای دکتر محمد رستمی، مرحوم سید علی قاضی، شهید چمران
سازمان	بانک، اداره‌کل، شرکت، بورس، وزارت، اداره، سازمان، اتحادیه، کنگره، کانون، مجمع، ارتش، هواپیمائی، حزب، امپراتوری، قرارگاه، کنسولگری، فرمانداری، استانداری، بخشداری، شهرداری، راهداری، سفارت، تیم، انتشارات، دولت، پارلمان، بنیاد، پاسگاه، کلانتری،	استانداری تهران، سفارت کشور فرانسه ، بنیاد سعدی، کانون دانش آموختگان دانشگاه فردوسی مشهد ، دانشکده سازمان صداسیما ، به گزارش گروه سیاست خارجی خبرگزاری فارس، تویوتا ، اتومبیل ایرانی، مسیحیان مقیم تهران، ارتش ایران، خبرگزاری
	گروه، ملیت، جامعه اقلیتهای مذهبی، شبکه، خبرگزاری، دانشگاه، فرهنگستان، دانشکده، کالج، آموزشکده، مدرسه، کودکستان، مهدکودک، کتابخانه، موسسه، انجمن، فدراسیون، کنفدراسیون، بیمارستان، مطب، کلینیک، فرودگاه، راه‌آهن، مرکز تحقیقاتی، حوزه علمیه، موسسه فرهنگی	فارس، نیروی هوایی آمریکا، بی بی سی، مجلس شورای اسلامی، گروه جی ۸، مرکز تحقیقاتی فیزیک نظری، دانشگاه تهران، کارگاه تولیدی کیف و کفش لارستان، فرودگاه بین المللی کابل
مکان	آتشفشان، دریاچه، دریاچه‌ی، ایالت، شهرک، روستا، ناحیه، شهر، استان، منطقه، رشته کوه، اقیانوس، دریای، دشت، کوه، خلیج، جزیره، صحرا، آبشار، بندر، جنگل، رود، رودخانه، چشمه، کوهستان، قاره، شهرستان، قاره، ستارگان، تنگه، خوابگاه، مترو، ترمینال، پایانه، پاساژ، پارک، بوستان، موزه، استادیوم، ورزشگاه، باشگاه، زورخانه، استخر، بدنسازی، باغ‌وحش، باغ، هتل، متل، تئاتر، رستوران، برج، تالار، پارکینگ، ایستگاه، جاده، بزرگراه، اتوبان، آزادراه، چهارراه، کوچه، بن‌بست، ساختمان، کاروانسرای، پل، زیرگذر، روگذر، تونل، سد، میدان، خیابان، کاخ، نیروگاه، سالن، آمفی‌تئاتر، نمایشگاه، مسجد، حرم، حسینیه ، امامزاده، کلیسای، معدن، کارخانه، مجموعه ورزشی، سرای محله، گاراژ	دریای خزر، جنگل سه هزار تنکابن، رشته کوه آلپ،شمال آفریقا، تنگه هرمز، مجموعه ورزشی آزادی ، کاخ گلستان، خیابان ۳۵ متری امیر کبیر، پارک آب و آتش، سد طالقان، کاروانسرای شاه عباسی کرج، شهرک گلستان، پاساژ ولیعصر، تونل کندوان، پل طبیعت، امامزاده داود کن، برج ایفل، شهرک صنعتی طوس، مترو ۴ راه ولیعصر، نمایشگاه عفاف و حجاب، سالن بدنسازی قهرمانان، آمفی تئاتر دانشکده روانشناسی دانشگاه تهران، مرکز همایشهای بین المللی برج میلاد، ساختمان میعاد

۳-۶- الگوهای متنی

به‌طورمعمول در متن‌های فارسی درصورت مشاهدهٔ برخی از الگوها به موجودیت‌های نامدار می‌توان دست پیدا کرد. این الگوها اغلب به‌صورت دنباله‌ای از کلمات که بعضی از واژه‌های آن ثابت هستند، پدیدار می‌شوند. همچنین برخی از الگوها نیز از دستهٔ عبارات باقاعده هستند که برای تشخیص

رایانامه، تاریخ و یا کمیت‌های عددی از آن‌ها می‌توان استفاده کرد. در این نوع روش‌ها باید به دنبال الگوها یا عباراتی گشت که اگر در یک متن تکرار شوند، درون خود موجودیت‌های نامدار مختلفی را در بر می‌گیرند. برای جمع‌آوری این الگوها نیاز به شناسایی عبارات مختلفی که با موجودیت نامدار مطرح شده‌اند، است و سپس پربسامدترین آن‌ها را می‌توان استخراج کرد. نمونه‌ای از این قواعد را در زیر مشاهده می‌کنید:

سفر (موجودیت شخص) به (موجودیت مکان)

مثال: سفر حسن روحانی به نیویورک

جاده (موجودیت مکان) به (مکان)

مثال: جاده تهران به مشهد

(ابتدای جمله-موجودیت شخص) گفت، افزود، بیان‌کرد: البته برای این قاعده باید صافی‌های گوناگونی را در نظر گرفت، از جمله اینکه نباید فاصله بین فعل "گفت" و ابتدای جمله زیاد باشد. دوم‌این‌که کلمه تشخیص داده‌شده به‌عنوان موجودیت نامدار، نباید برچسب ضمیر داشته باشد.

مثال: حسن روحانی گفت:

کشورهایی (شهرهایی-افرادی-اشخاصی-کسانی) مانند (موجودیت مکان)،(موجودیت مکان) و ...

به گفته (موجودیت شخص)

به اتفاق (موجودیت شخص)

(موجودیت مکان) پایتخت/مرکز (موجودیت مکان)

در دیدار (موجودیت شخص) با (موجودیت شخص)

البته لازم است به این نکته اشاره شود که الگوها و قواعد زیادی برای تشخیص موجودیت‌های نامدار وجود دارد و نمی‌توان تمامی قواعد را بیان کرد.

۷- بهبود با استفاده از کسرهٔ اضافه

کسرهٔ اضافه یکی از مهم‌ترین ویژگی‌های هر کلمه در زبان فارسی است. در نوشتار زبان فارسی کسره اضافه واکه‌ای است که بیشتر در دو ترکیب موصوف و صفت و مضاف و مضاف‌الیه وجود دارد. کسرهٔ اضافه بخش بسیار مهمی از هر کلمه است که در ساختار نحوی و حتی معنایی، نقش مهمی را ایفا می‌کند. با این وجود در متون فارسی اگر کلمه به واکه ختم شود، کسرهٔ اضافه نمایش داده نمی‌شود و این عدم نمایش باعث ابهام در ساختار نحوی و معنایی جملات می‌شود. کسرهٔ اضافه در متن در تمامی لایه‌های صرف، نحو و معنا تأثیرگذار بوده و تشخیص آن به حل برخی چالش‌های موجود در پردازش متون زبان فارسی کمک خواهد کرد. کسرهٔ اضافه نقش کلیدی و اساسی را در شناسایی دامنهٔ موجودیت‌های نامدار ایفا می‌کند. سامانه‌های شناسایی موجودیت‌های نامدار در شناسایی دامنه آن‌ها ممکن است، تشخیص نادرست بدهند و با استفاده از کسرهٔ اضافه این چالش تا حدودی برطرف می‌شود. بسیاری از شاخص‌هایی که برای تشخیص موجودیت‌های نامدار مورد استفاده قرار می‌گیرد، باید به‌حتم همراه کسرهٔ اضافه باشند(مانند آقای، خانم، سازمان، سالن و غیره). همچنین انتهای دامنهٔ یک موجودیت نامدار نیز با استفاده از کسرهٔ اضافه با دقت بیشتری شناسایی می‌شود. در مورد موجودیت شخص بیش‌تر شاخص‌ها بدون کسرهٔ اضافه هستند(مانند سردار، سرکار، دکتر، مهندس، مرحوم، ژنرال و غیره) اما در مورد سایر موجودیت‌های نامدار، بیشتر شاخص‌ها باید همراه با کسرهٔ اضافه بیایند. به‌عنوان مثال شاخص‌های حضرت، جناب، آقای، خانم باید دارای کسره بوده تا به همراه کلمات بعد از آن یک موجودیت نامدار را شامل شوند و برای بعضی از شاخص‌ها مانند سردار، سرتیپ، دکتر، مهندس و نظیر آن نیازی به وجود کسره اضافه نیست. نکته دیگر در مورد ادامه یک موجودیت نامدار است. در این الگو تا زمانی‌که کلمه دارای کسرهٔ اضافه است به‌عنوان موجودیت نامدار در نظر گرفته می‌شود. به‌عنوان نمونه عبارت "مجموعهِ ورزشیِ آقایِ سعیدِ قهرمانیِ لنگرودی" بیانگر یک موجودیت نامدار است و این شناسایی به واسطهٔ استفاده از کسرهٔ اضافه در سامانه است. برای نمونه در جمله دولت روحانی اعلام کرد عبارت "دولت روحانی اعلام" به‌عنوان موجودیت نامدار سازمان شناسایی شده است که اگر کسرهٔ اضافه در این عبارت دخیل شود، فقط عبارت "دولت روحانی" به‌عنوان موجودیت نامدار شناسایی می‌شود. با استفاده از کسرهٔ اضافه در قوانین و همچنین بهبود دامنه یک موجودیت نامدار به‌وسیلهٔ کسرهٔ اضافه، دوباره فرآیند ارزیابی تکرار شده و در قسمت نتایج به آن پرداخته خواهد شد.

۸- ارزیابی

با توجه به عدم وجود دادهٔ آزمون برای تشخیص موجودیت نامدار در زبان فارسی و اهمیت وجود آن، دادهٔ آزمون موجودیت نامدار ایجاد شد و با استفاده از داده‌های پیکرهٔ بی جن‌خان، به‌طور دستی توسط افراد خبره انجام و سه نوع

شخص، سازمان و مکان برای هر کلمه برچسب‌زنی شد[1]. کلماتی که بیان‌گر موجودیت نامدار نباشند با برچسب O مشخص شده‌اند. علت انتخاب مجموعه آزمون از پیکرهٔ بی‌جن‌خان [3] استانداردبودن متن و پراکندگی انواع گوناگون متون است. همچنین با توجه به اینکه برای سامانه تشخیص موجودیت نامدار استفاده از برچسب اجزای سخن اهمیت زیادی دارد؛ لذا باید داده‌ای که برای آزمون انتخاب می‌شود از برچسب اجزای سخن با دقت صد درصد داشته باشد تا سامانه را به‌صورت درست بتوان ارزیابی کرد. حجم داده آزمون ۴۲۰۰۰ کلمه است که از این تعداد ۲۳۰۴ کلمه، موجودیت نامدار هستند. پراکندگی موجودیت‌های نامدار به‌صورت جدول (۳) است:

(جدول-۳): پراکندگی انواع موجودیت نامدار در داده آزمون
(Tabel-3): Distribution of Named Entity in testSet

	سازمان	شخص	مکان
تعداد کلمه	۶۸۲	۴۶۶	۱۱۵۶

جهت ارزیابی سامانه تشخیص موجودیت نامدار از یک مدل امتیازدهی که برای ارزشیابی در کنفرانس MUC[2] مطرح شده، استفاده شد [9]. این ارزشیابی از دو معیار دقت[3] (P) و بازخوانی[4](R) انجام می‌گیرد. این معیارها در حقیقت از تعاریف موجود در حوزهٔ بازیابی اطلاعات اخذ شده‌اند. جهت محاسبهٔ آن‌ها پس از اجرای آزمون و مقایسهٔ خروجی با داده‌های اصلی جدول (۴) به‌دست می‌آید. به این جدول به‌اصطلاح ماتریس ابهام[5] می‌گویند.

$$P=TP/(TP+FP) \quad (1)$$

$$R=TP/(TP+FN) \quad (2)$$

$$F=2*P*R/(P+R) \quad (3)$$

برای محاسبه معیارهای TP,FN,FP,FP برچسب تک‌تک کلمات مقایسه می‌شود. اگر سامانه کلمه‌ای را به‌عنوان موجودیت نامدار شناسایی کرده بود و در داده اصلی هم آن کلمه موجودیت نامدار بود TP یک واحد افزایش می‌یابد. اگر سامانه کلمه‌ای را به‌عنوان موجودیت نامدار تشخیص داد، اما در اصل آن کلمه موجودیت نامدار نبود FP یک واحد افزایش پیدا می‌کند. اگر کلمه‌ای را به‌عنوان موجودیت نامدار شناسایی نکرده بود و در داده اصلی هم موجودیت نامدار نبود TN یک واحد افزایش پیدا می‌کند. اگر هم سامانه کلمه‌ای را به‌عنوان موجودیت نامدار شناسایی نکرده بود، ولی درواقع آن کلمه موجودیت نامدار بود FN یک واحد افزایش پیدا می‌کند.

(جدول-۴): ماتریس ابهام
(Tabel-4): Confusion Matrix

		تشخیص سیستم	
		مثبت	منفی
کلید (پاسخ‌های صحیح)	مثبت	TP[7]	FN[6]
	منفی	FP[9]	TN[8]

از ترکیب معیارهای فراخوانی و دقت، معیار دیگری به دست می‌آید که بیان‌گر میانگین همساز این دو معیار است و به آن معیار F یا F-Measure گویند.

۹- یافته‌ها

یافته‌های حاصل از ارزیابی برای سه نوع مکان، شخص و سازمان به همراه ارزیابی کلی در جدول (۵) بیان شده است.

(جدول-۵): نتایج سامانه تشخیص موجودیت نامدار
(Tabel-5): Ner Results

	P(درصد)	R(درصد)	F(درصد)
مکان	۸۴	۸۶	۸۴/۹۸
شخص	۸۶	۶۷	۷۵/۳۲
سازمان	۶۹	۷۵	۷۱/۸۷
کل	۷۵	۸۳	**۷۸/۷۹**

سپس در قواعد تولیدشده کسرهٔ اضافه به‌عنوان یکی از مهمترین ویژگی‌های زبان فارسی اضافه و سپس دوباره فرآیند ارزیابی تکرار شد و نتایج به‌صورت جدول (۶) قابل مشاهده است.

[1] داده آزمون:
http://cdn.persiangig.com/preview/rplIAZAgPn/large/input3TypeNERTestCase.zip

[2] Message Understanding Conference

[3] Precision

[4] Recall

[5] Confusion Matrix

[6] False Negative

[7] True Positive

[8] True Negative

[9] False Positive

(جدول-۶): نتایج سامانه تشخیص موجودیت نامدار با استفاده از کسرهٔ اضافه

(Tabel-6): Ner Results using Ezafe

	P(درصد)	R(درصد)	F(درصد)
مکان	۹۱	۸۶	**۸۸/۴۲**
شخص	۸۶	۶۸	**۷۵/۹۴**
سازمان	۸۰	۷۴	**۷۶/۸۸**
کل	۸۳	۸۱	**۸۱/۹۸**

همانگونه که ملاحظه می‌شود، سامانه بهبود بیشتری در انواع موجودیت مکان و سازمان داشته است. علت این امر وجود تنوع بیشتر برچسب‌های اجزای سخن در انواع مکان و سازمان نسبت به شخص است. به‌عنوان نمونه در یک موجودیت نامدار مکان یا سازمان برچسب‌های اجزای سخن حرف‌ربط، قید، صفت و اسم ممکن است وجود داشته باشد؛ اما تنها برچسب‌های اسم یا صفت در موجودیت شخص حضور می‌توانند داشته باشند. با توجه به تنوع بیشتر کلمات و برچسب‌های مختلف در موجودیت‌های مکان و سازمان نسبت به شخص و تشخیص درست انتهای دامنه موجودیت‌های مکان و سازمان با استفاده از کسرهٔ اضافه نتایج مذکور قابل توجیه است. نتایج بیان‌گر آن است که سامانهٔ پیشنهادی در مقایسه با سامانهٔ مشابه تشخیص موجودیت نامدار در زبان فارسی که مبتنی بر قاعده توسعه داده شده‌اند، عملکرد بهتری دارد. جدول ۷ مقایسه بین سامانه‌های تشخیص موجودیت نامدار را در زبان فارسی نشان می‌دهد.

(جدول-۷): مقایسه سامانه‌های تشخیص موجودیت نامدار زبان فارسی

(Tabel-7): Result Comparison with similar system

ابزارهای تشخیص موجودیت نامدار	P(درصد)	R(درصد)	F(درصد)
ابزار مرتضوی و همکاران	۷۲	۷۶	۷۳/۹۴
ابزار پیشنهادی	**۸۳**	**۸۱**	**۸۱/۹۸**

۱۰- نتیجه‌گیری

سامانهٔ پیشنهادی بر اساس روش مبتنی بر قاعده به‌همراه فهرست واژگان اقدام به شناسایی موجودیت‌های نامدار می‌کند. استخراج واژگان از ویکی‌پدیای فارسی که حجم عظیمی از موجودیت‌های نامدار را شامل می‌شود، صورت پذیرفت. استخراج این اسامی از ویکی‌پدیا به‌خاطر حجم عظیم، چالش‌هایی را به همراه داشت که پیش‌پردازش‌هایی جهت رفع آن‌ها اعمال شد. همچنین برای دسته‌بندی اسامی ویکی‌پدیا از رده‌های منتسب به هر عنوان استفاده شد. بعد از استخراج عناوین رده‌ها نیز فهرست موجودیت‌های نامدار به‌منظور رفع ابهام‌های احتمالی به‌صورت دستی بررسی شد. موضوع مطروحهٔ دیگر در مورد اسامی جدیدی است که در منبع واژگان ویکی‌پدیا وجود نداشت. با اعمال قواعد ریخت‌شناسی، هم‌جواری و الگوهای متنی، این موجودیت‌ها مورد شناسایی قرار گرفت. درنهایت برای ارزیابی سامانه، ۴۲ هزار توکن از پیکرهٔ بی‌جن‌خان به‌صورت دستی برچسب‌زنی شد. علت انتخاب پیکرهٔ بی‌جن‌خان استاندارد و درست‌بودن کلمات و برچسب اجزای سخن آن‌ها بود. معیار F اولیه ۷۸/۷۹ است که بدون استفاده از کسرهٔ اضافه در قواعد به‌دست آمد. در ادامه برای بهبود در شناسایی موجودیت‌های نامدار و دامنه آن‌ها از کسرهٔ اضافه استفاده شد. کسرهٔ اضافه یکی از ویژگی‌های مهم و پنهان کلمات زبان فارسی است؛ سپس دوباره فرآیند ارزیابی تکرار و معیار F به ۸۱/۹۸درصد افزایش یافت. نتایج بیانگر بهبود چهار درصدی معیار F با استفاده از کسرهٔ اضافه است. این سامانه در مقابل سامانهٔ قاعده‌بنیان مشابه فارسی عملکرد بهتری را از خود نشان می‌دهد.

۱۱- پیشنهادها

با توجه به اهمیت روش‌های یادگیری ماشین برای تشخیص موجودیت‌های نامدار پیشنهاد می‌شود با استفاده از ابزار فعلی، پیکره‌ای را تولید و از آن برای توسعه روش‌های یادگیری ماشین بهره برد. بهترین راه برای تولید پیکره این است که از داده‌های استانداردی مانند پیکره بیجن‌خان استفاده کرد و برچسب‌زنی اولیه را با استفاده از قواعد بر روی آن انجام داد؛ سپس به‌صورت دستی برچسب‌ها اصلاح شوند تا داده آموزش تشخیص موجودیت نامدار تولید شود؛ همچنین برای افزایش دقت دامنه موجودیت نامدار از تجزیه نحوی می‌توان استفاده کرد. از پیشنهادهای دیگری که منجر به افزایش دقت تشخیص موجودیت نامدار می‌شود، سامانهٔ تشخیص موجودیت نامدار بر اساس روش یادگیری ماشین در کنار روش‌های باقاعده و مبتنی بر واژگان است.

۱۲- مراجع 12-References

[۱] اصفهانی سیدعبدالحمید, راحتی قوچانی سعید, جهانگیری نادر.«سیستم شناسایی و طبقه‌بندی اسامی در متون فارسی». فصلنامه پردازش علایم و داده‌ها. شماره ۱۳. ۷۷-۷۸. ۱۳۸۹

[1] Esfahani.A, Rahati.S, Jahangiri.N. "Identification and classification names in Persian texts ." Signal and Data Processing Journal ,No 13,78-77, 1389

[۲] سادات مرتضوی پونه و شمس‌فرد مهرنوش. «شناسایی موجودیت‌های نامدار در متون فارسی». پانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران. تهران. انجمن کامپیوتر. مرکز توسعه فناوری نیرو. ۱۳۸۸

[2] Mortazavi.P, Shamsfard.M."Named Entity Recognition In Persian Texts". 15nd National Computer Society of Iran Conference.tehran. Power Technology Development Center.Tehran. 1388

[3] Bijankhan.M, Sheykhzadegan.J, Bahrani.M and Ghayoomi.M. "Lessons from Building a Persian Written Corpus:Peykare." Language Resources and Evaluation.2011. pp. 143-164.

[4] Chieu, Hai Leong, and Hwee Tou Ng. "Named entity recognition: a maximum entropy approach using global information." Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002.

[5] Das, Arjun, Debasis Ganguly, and Utpal Garain. "Named Entity Recognition with Word Embeddings and Wikipedia Categories for a Low-Resource Language." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 16.3 (2017): 18.

[6] Elsebai, Ali. "Arabic Proper Names Recognition Using Heuristics." Proceeding of the 9th Annual Post Graduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNET), ISBN. 2008.

[7] B. Farber, D. Freitag et al."Improving NER in Arabic Using a Morphological Tageer". the 6th International Conference on Language Resources and Evaluation,LREC. 2008.

[8] Farmakiotou, Dimitra, et al. "Rule-based named entity recognition for Greek financial texts." Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000). 2000.

[9] Grishman R, Sundheim B." Message Understanding Conference-6: A Brief History". InCOLING 1996 Aug 5 (Vol. 96, pp. 466-471).1996

[10] Mansouri, Alireza, Lilly Suriani Affendey, and Ali Mamat. "Named entity recognition approaches." International Journal of Computer Science and Network Security 8.2: 339-344. 2008

[11] Mikheev, Andrei, Marc Moens, and Claire Grover. "Named entity recognition without gazetteers." Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1999.

[12] Rau, Lisa F. "Extracting company names from text." Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on. Vol. 1. IEEE, 1991.

[13] Shaalan, Khaled, and Hafsa Raza. "Person name entity recognition for Arabic." Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources. Association for Computational Linguistics, 2007.

[14] Tjong Kim Sang, Erik F., and Fien De Meulder. "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.

محمد عبدوس تحصیلات کارشناسی مهندسی نرم افزار را درسال ۱۳۹۲و کارشناسی ارشد خود را در گرایش هوش مصنوعی، سال ۱۳۹۴ در دانشگاه علم و صنعت ایران به پایان رسانده و در حوزه‌های یادگیری ماشین، پردازش زبان طبیعی، متن‌کاوی و داده کاوی به فعالیت پرداخته است.

نشانی رایانامه ایشان عبارت است از :

mohammadabdous@comp.iust.ac.ir
md.abdous@gmail.com

بهروز مینایی بیدگلی دکترای خود را در رشته علوم و مهندسی کامپیوتر از دانشگاه ایالتی میشیگان آمریکا در سال ۱۳۸۴ گرفت. تخصص وی نرم‌افزار، هوش مصنوعی، پردازش زبان طبیعی و داده‌کاوی است. ایشان هم‌اکنون به‌عنوان عضو هیئت علمی دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت به تدریس دروس مختلف نرم‌افزار و هوش مصنوعی مشغول است.

نشانی رایانامه ایشان عبارت است از:

B_minaei@iust.ac.ir