شناسایی موجودیت‌های نام‌دار در متون فارسی

پونه سادات مرتضوی
آزمایشگاه تحقیقاتی پردازش زبان طبیعی، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران
pooneh_mortazavi@yahoo.com

مهرنوش شمس فرد
آزمایشگاه تحقیقاتی پردازش زبان طبیعی، دانشکده مهندسی برق و کامپیوتر،دانشگاه شهید بهشتی، تهران، ایران
m-shams@sbu.ac.ir

چکیده: شناسایی موجودیت‌های نامدار[1] در پردازش زبان طبیعی به عملیاتی گفته می‌شود که در طی آن کلیه‌ی اسامی خاص موجود در متن و متعلق به مقوله‌های معنایی مختلف، شناسایی و استخراج می‌گردند.

در این مقاله به معرفی سیستمی توسعه یافته به منظور تشخیص اسامی نامدار و دسته‌بندی آنها در زبان فارسی پرداخته می‌شود. این سیستم با بکارگیری ساختار واژوی اسامی خاص و نیز الگوهای متنی ممکن برای اسم‌های خاص متعلق به یک دسته، سعی در شناسایی موجودیت‌های نامدار می‌کند. علاوه بر آن، با بکارگیری برچسب نحوی و معنایی برای هر کلمه و توجه به سایر رخدادهای آن در متن، عملیات رفع ابهام برای بهبود شناسایی را انجام می دهد.

واژه های کلیدی: شناسایی موجودیت‌های نامدار، پردازش زبان طبیعی، مقوله بندی موجودیت‌های نامدار، برچسب معنایی، الگو شناسایی، ریخت شناسی موجودیت های نامدار.

۱- مقدمه

موجودیت نامدار به کلمه و یا عبارتی گفته می‌شود که برای ارجاع به نمونه های یک مقوله‌ی مشخص مانند شخص، شرکت یا موسسه، تاریخ، بیماری، گونه‌ای باکتری و سایر بکار می‌رود. نیاز به شناسایی موجودیت‌های نامدار، در دنیای امروز که عصر ارتباطات و اطلاعات است رو به رشد می‌باشد. شناسایی موجودیت‌های نامدار برای جستجوهای معنادار، ترجمه‌ی خودکار، استخراج خودکار مفاهیم متن، کشف ارجاعات درمتن و بسیاری دیگر از زمینه‌های مربوط به پردازش زبان های طبیعی کاربرد دارد.

اینکه سیستم چه نوع موجودیتی را تشخیص دهد و یا به بیان دیگر دسته های معنایی موردنظرش چه باشند، وابسته به زمینه‌ی کاربردی سیستم می‌باشد. شناسایی موجودیت نامدار در علم زیست شناسی می‌تواند تشخیص اسامی وابسته به انواع «پروتئین‌ها»، «DNA»ها، «نوع سلول» و ... ، در حوزه پزشکی می تواند تشخیص انواع بیماری ها، داروها، مراکز درمانی و مانند این ها و در حوزه تجارت نام شرکت ها و موسسات، تراکنش های مالی، بورس و غیره باشد. همچنین این امر می تواند به صورت خیلی خاص مثلاً فقط برای کشف اسامی شرکت‌های تولید کننده فولاد از روی متون مربوطه بکار رود. یک دسته بندی عام که در بسیاری تحقیقات NER مورد استفاده قرار می گیرد و در MUC-6[۲] نیز ملاک مقایسه بوده است، دسته بندی معنایی بر اساس «شخص»، «سازمان» و «ناحیه» است. در این مقاله پس از مروری بر فعالیت های انجام شده در این حوزه به معرفی سیستم NER پیاده سازی شده برای زبان فارسی می پردازیم. در این معرفی ابتدا به دسته های معنایی مورد نظر در این سیستم اشاره نموده سپس با ارائه معماری کلی سیستم، رهیافت مورد استفاده در آن را شرح خواهیم داد.

۲- کارهای مرتبط

در زبان‌های مختلف بر روی مبحث «تشخیص و دسته بندی اسامی نامدار (NERC)» به عنوان زیر شاخه‌ای از پردازش زبان طبیعی کارهایی انجام شده است. اگرچه نسبت بالایی از تحقیقات NERC متعلق به زبان انگلیسی بوده است، فعالیت هایی برای زبان‌هایی چون آلمانی، یونانی، ژاپنی، فرانسوی، ایتالیایی، هلندی، چینی، روسی، کره‌ای، رومانیایی و ترکی نیز انجام شده است. اکثر کارهای انجام شده متعلق به زبان، دامنه و یا گونه نوشتاری خاص است و بزرگ‌ترین مشکل این گونه سیستم‌ها نیز مربوط به انتقالشان به دامنه‌ی جدید می باشد.

در این راستا برای زبان فارسی، کار قابل توجهی انجام نگرفته است؛ البته در مواردی در حواشی موضوع کارهایی شده است که از جمله می توان به تعیین عبارات اسمی هم-مرجع در فارسی[۱] اشاره نمود.

روش‌های اولیه‌ای که برای کشف موجودیت‌های نامدار پیشنهاد می‌شد بیشتر مبتنی بر قانون بود، در حالی که امروزه اکثر کارها به سمت استفاده از روش‌های یادگیری پیش می‌رود [۳] و [۴].

سیستم‌های توسعه یافته بر پایه یادگیری برای NERC به سه دسته تقسیم می‌شوند[۵]: یادگیری با نظارت[۲](SL)، یادگیری نیمه‌نظارت[۳](SSL) و یادگیری بی‌نظارت[۴](UL). در یادگیری با نظارت، نیاز به حجم زیادی از متون حاشیه نویسی شده است و از تکنیک‌هایی نظیر درخت تصمیم‌گیری و مدل مخفی مارکوف (HMM) استفاده می‌کند. این روش مناسب برای سیستم‌های وابسته به دامنه است (۳). در یادگیری نیمه‌نظارت از مجموعه‌ای از کلمات نمونه برای یادگیری استفاده می‌شود (مثل مجموعه‌ای از نام کتاب‌ها). این روش ها ابتدا سعی در یافتن موجودیت‌های نمونه در متون کرده و سپس با توجه به ساختار متن در بر دارنده کلمات، بدنبال سایر موجودیت‌هایی از آن نوع می‌گردد. یادگیری بدون نظارت، با استفاده از منابع لغوی چون WordNet[۵] و الگوهای لغوی و محاسبات آماری به شناسایی می‌پردازد. به عبارتی تکنیک معمول برای یادگیری بدون نظارت بر پایه دسته بندی موجودیت‌ها است[۶].

در زبان فارسی با توجه به عدم وجود پیکره های برچسب خورده با برچسب های معنایی یا طبقات نامدار و همچنین دردسترس نبودن واژگان های کاملی چون وردنت به ناچار باید تمرکز بیشتر بر روش های مبتنی بر قاعده باشد. سیستم معرفی شده در این مقاله از لحاظ استفاده از قواعد زبان شناسی، مشابه سیستم توسعه یافته [۷] عمل می‌کند با این تفاوت که در زبان فارسی همچون زبان انگلیسی امکان درک موجودیت نامدار با توجه به بزرگی حرف اول کلمات وجود ندارد. پس بجای آن برای بهبود سیستم از برچسب‌گذاری‌های نحوی و مجموعه‌ای از اسامی خاص مانند مجموعه اسامی مناطق، افراد، روزهای هفته، واحدهای پولی، عناوین و... استفاده می‌شود. استفاده از ریخت-شناسی کلمات از دیگر نقاط قوت این سیستم می‌باشد که در مقاله [۸] هم به نوعی به آن اشاره شده است.

۳- معماری سیستم پیشنهادی

سیستم NER معرفی شده در این مقاله از سه بخش اصلی (۱) پیش پردازش، (۲) تشخیص و برچسب زنی و (۳) پالایش و تولید نتیجه تشکیل شده است. شمای کلی از معماری سیستم در شکل (۱) نمایش داده شده است.

در این سیستم پس از دریافت متن، با ورودی مورد پذیرش سیستم تطبیق داده می‌شود، واحدهای خاص شناسایی شده و سپس بر روی آن برچسب POS زده می‌شود. بعد از این برچسب گذاری نحوی، با توجه به ترکیب‌های اعداد، کلمات و علامات به شناسایی اولیه اعداد و برچسب گذاری‌های معنایی پرداخته می-شود و برچسب گذاری ثانویه‌ای از نام‌های خاص مکمل این مرحله است. در انتهای مرحله پیش پردازش، با استفاده از ساختار واژگانی اسامی خاص، سعی در شناخت این واژگان مستقل از متن می‌شود.

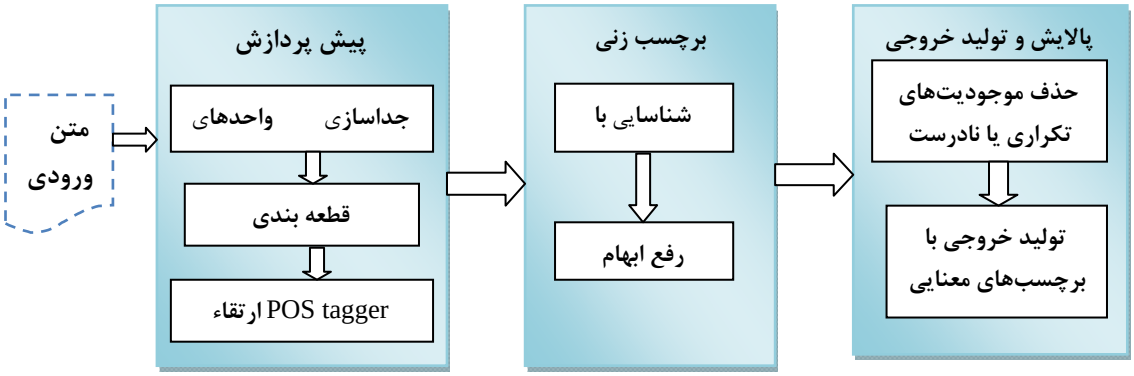
در مرحله‌ی شناسایی، با توجه به نحوه‌ی چیدمان واژگان در متن و الگوهای موجود برای هر مقوله، موجودیت‌های نامدار شناخته شده و مرز آنها با توجه به تکرارها و برچسب نحوی و معنایی کلمات مشخص می‌شود.

در انتها با حذف خطاها و تکرارهای شناخته شده، هر موجودیت در مقوله‌ی مناسب خود دسته بندی شده و خروجی با برچسب گذاری معنایی موجودیت‌های نامدار با عنوان مقوله‌ای که در آن قرار می‌گیرند تولید می‌شود.

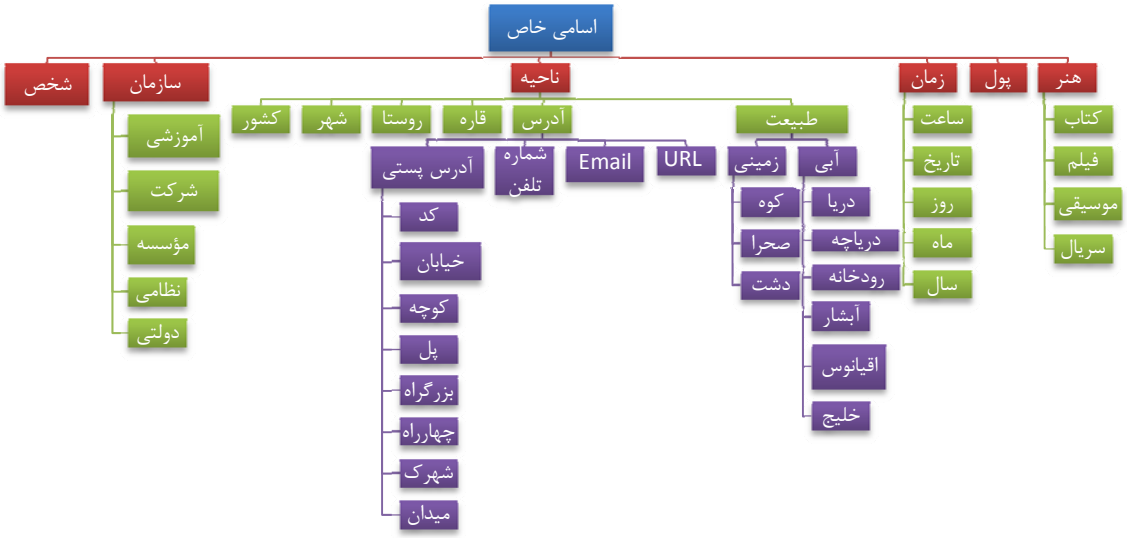
در ادامه پس از معرفی مجموعه برچسب های معنایی مورد استفاده در این سیستم به ارائه نحوه عملکرد سیستم در بخش های مجزا خواهیم پرداخت.

۴- مقوله بندی برچسب های معنایی

موجودیت‌های نامدار در مقوله‌های مختلفی دسته بندی می‌شوند. این مقوله‌ها می‌توانند بسته به هدف سیستم تشخیص اسامی نامدار، در طیفی از زمینه‌های تخصصی چون تشخیص اسامی بیماری‌ها و «DNA» و ... تا زمینه‌های عامی چون نام افراد یا سازمان‌ها رده بندی شوند. همچنین در این طبقه بندی هر مقوله‌ی کلی هم می‌تواند به زیر مقوله‌های جزئی‌تر تقسیم شود.



شکل ۱: شمای کلی از معماری سیستم



شکل ۲: بخشی از سلسله مراتب برچسب های معنایی

در این سیستم، ساختار سلسله مراتبی برای مقوله‌های موجودیت‌های خاص در دسته‌ی عام طراحی شده است که در شکل (۲) به تصویر کشیده شده است.

یک موجودیت نامدار می‌تواند در هر رده از سلسله مراتب بالا قرار گیرد. هر چه در عمق پایین‌تر، برچسب معنایی آن دقیق‌تر است.

۵- شناسایی موجودیت‌های نامدار در جمله فارسی

تشخیص موجودیت‌های نامدار به طور کلی به دو صورت انجام می‌شود، یکی به صورت درونی که با استفاده از برچسب-های معنایی و برچسبهای مقوله نحوی نسبت داده شده به هر کلمه انجام می‌گیرد و شیوه‌ی دوم به صورت برونی است و تشخیص موجودیت نامدار با توجه به متنی که کلمه در آن قرار گرفته است و کلمات اطراف آن صورت می‌گیرد.

این سیستم، از هر دو روش استفاده می‌کند و در هر کدام برای افزایش صحت جواب حاصله، از چندین راه حل ممکن بهره می‌گیرد. در ادامه به بیان این راه حل‌ها و نحوه‌ی بکارگیری آنها در سیستم پرداخته می‌شود.

۵-۱- پیش‌پردازش

پس از دریافت متن ورودی و پیش از انجام هر گونه عملیاتی بر روی متن، ابتدا واحدهای خاص (چون URL) شناسایی می-شوند و سپس پردازش‌های اولیه مورد نیاز برروی متن، اجرا می-شوند. این پردازش ها شامل قطعه بندی و تحلیل ساختواژی احتمالی جهت آماده سازی متن برای ورود به پیمانه برچسب زن مقوله نحوی ارتقاء یافته است.

در این پیمانه پس از برچسب زنی نحوی معمولی، برای اصلاح برچسب‌های نحوی ایجاد شده برای هر کلمه، از نوع دیگری از برچسب‌های اسم خاص ذخیره شده در پایگاه داده استفاده می‌شود. همچنین برچسب‌های معنایی اولیه، برای اعداد و ترکیب آنها با علامات اطراف و نیز ساخت شناسی کلمات خاص، در این مرحله زده می‌شود تا در مرحله تشخیص، اصلاح و تکمیل شوند.

۵-۱-۱- استخراج واحدهای خاص

برخی از واحدهای متنی لازم است به همان شکل وارد شده در متن و بدون هیچ پردازش یا تغییر شکلی استخراج شوند. از جمله این واحدها می توان به آدرس صفحات وب (URL) و آدرس نامه‌های الکترونیکی اشاره نمود. این آدرس ها لازم است قبل از انجام قطعه بندی[۶] متن جدا شوند چراکه بدلیل پیش پردازش‌هایی که بر روی متن ورودی صورت می گیرد و ایجاد فواصل میان علائم نگارشی با سایر کلمات، یک آدرس URL و یا Email که دنباله‌ی چسبیده‌ی حروف در آنها حائز اهمیت است، از هم جدا شده و درصد درستی تشخیص را پایین می‌آورد. به همین منظور کشف این دو موجودیت نامدار که جزو دسته آدرس در ساختار سلسله مراتبی برچسب های معنایی قرار می-گیرند، باید پیش از این جداسازی‌ها انجام گیرد. پس از جمله اولین اقداماتی که انجام می‌گیرد شناسایی این دو مقوله در متن و جداسازی آنها پس از شناسایی از متن می‌باشد.

از جمله فواید جداسازی‌ها افزایش دقت در شناسایی و نیز افزایش سرعت در tagger و مراحل پس از آن می‌باشد. اما از

طرف دیگر، ورود آن به کد از ویژگی سادگی در ارتقاء آن می-کاهد.

پیش از انجام شناسایی باید فرمت نگارشی و علائم این دو آدرس بررسی شده و سپس با در نظر گرفتن آنها رشته‌ای را که دارای آن ویژگی‌ها است در یکی از مقوله‌های ذکر شده قرار داد و آن را از متن برای جلوگیری از پردازش‌های اضافی حذف کرد.

از آنجایی که مهترین کاراکتری که در شناسایی Email مؤثر است، قابل استفاده در URL نیز می‌باشد، مهم است که پیش از شناسایی Emailها به شناسایی URL پرداخته شود.

۵-۱-۱-۱- فرمت آدرس Email

برای تشخیص آدرس Email در متن باید دانست که از چه اجزایی تشکیل شده است و در هر جزء آن چه کاراکترهایی مجاز به استفاده می‌باشند. هر آدرس Email به طور کلی از دو جزء «بخش محلی[۷]» و «دامنه اینترنتی[۸]» تشکیل شده است و به این شکل می‌باشد:

«local-part "@" domain». «بخش محلی» که آغازگر آدرس است با علامت @ از «دامنه» جدا می‌شود. هر یک از این دو جزء هم فرمت و محدودیت‌های کاراکتری خود را دارد و گرامر صحیح آن در RFC آمده است. برای کشف صحیح Email باید از این گرامر پیروی کرد.[۹]

۵-۱-۱-۲- فرمت URL [۱۰]

URL به رشته‌ای از کاراکترها گفته می‌شود که بدون هیچ فاصله‌ای از یکدیگر آمده‌اند و نشان دهنده‌ی منبعی موجود در اینترنت می‌باشد. ویژگی‌های آن از مفاهیم بیان شده توسط درک اطلاعات سراسری شبکه گسترده جهانی مشتق شده است.

به طور کلی طبق استاندارد بیان شده در rfc1738 و rfc4395، URLها به این صورت نوشته می‌شوند: <بخش خاص شِما>:<شِما[۹]>

تفسیر «بخش خاص شِما» به «شِما» بستگی دارد. اسامی شِما، از دنباله‌ای از کاراکترها تشکیل می‌شود. کاراکترهای لاتین می‌توانند به صورت بزرگ هم نوشته شوند که مفسر آنها را به حروف کوچک تبدیل می‌کند. (مثلاً «HTTP» هم همچون «http» ممکن است.) شناسایی URL تنها به کاراکترهایی که برای آن استفاده شده است بستگی دارد. پس برای شناسایی URL تنها لازم است کاراکترهای مجاز برای استفاده در آن را بیابیم.

هر شِمای مورد استفاده در سطح اینترنت، پیش از معرفی در مخزنی در IANA[۱۰] ثبت می‌شود. به طور کلی تمامی URLها در IANA وجود دارند و هر شِمایی که جدید تعریف می‌شود باید علاوه بر نام، شیوه و علائم دسترسی به منابع درون آن شِما را معرفی کند.

شِماهای ثبت شده در IANA به سه دسته تقسیم می‌شوند: «شماهای دائمی»، «شماهای موقتی» و «شماهای قدیمی». اعضای هر یک از این سه دسته، طبق آخرین به روز رسانی مخزن IANA در پایگاه داده‌های سیستم ذخیره شده است و برای شناسایی URLها بکار می‌رود. [۱۱]

۵-۱-۲- قطعه بندی و برچسب‌گذاری مقوله نحوی ارتقاء یافته

پس از استخراج واحدهای خاص متن تحت عمل قطعه بندی قرار می گیرد تا مرز کلمات و جملات آن مشخص شود. برای عمل قطعه بندی، کلیه عملیات لازم اعم از جدا کردن کلمات چسبیده، برداشتن فاصله‌ها و نیم‌فاصله‌های اضافه، چسباندن کلمات جدا را سیستم انجام می‌دهد. بعد از این پردازش‌های اولیه، متن حاصل به برچسب زن مقوله نحوی[۱۱] داده می شود. این برچسب زن [۱۲] در آزمایشگاه NLP دانشگاه شهید بهشتی توسعه یافته و متن را بر اساس مجموعه‌ای از ۱۹ نوع برچسب[۱۲] برچسب‌گذاری می کند.

پس از برچسب گذاری نحوی، محل هر واژه در جمله هم به عنوان یک نمایه قابل بازیابی به کلمه الصاق و نگهداری می‌شود. این امر موجب می‌شود در صورت نیاز (در مراحل بعدی) بتوان به کلمات پس از آن در متن دسترسی داشت و از خاصیت تکرار واژه‌های پشت هم در یک متن بهره برد. همچنین در مرحله‌ی نهایی برای تشخیص و اصلاح خطاهایی نظیر شناسایی اسامی خاص موجود در اسم خاص دیگر، نیاز به نگهداری جای هر واژه در متن است.

علاوه بر آن حروفی را که تحلیلگر ساختواژی مورد استفاده از کلمات حذف کرده است (مثل «ها» یا ضمیرهای متصلی چون «م») برای هر کلمه نگهداری می‌شود تا در مراحل آتی از آنها استفاده شود. این حروف خود می‌توانند عامل تشخیص دهنده‌ی عدم تعلق یک واژه به یک نوع موجودیت باشند.

۵-۱-۲-۱- برچسب زنی اعداد

در سیستم شناسایی موجودیت های نامدار، علاوه بر برچسب‌زنی معمول نیاز به برچسب های پیشرفته تری نیز هست

که بر اساس ترکیب برچسب های قبلی بدست می‌آیند. یک دسته از این برچسب ها مربوط به اعداد صحیح و اعشاری نوشته شده به صورت حرفی و عددی و همچنین نمایش های مختلف زمان و تاریخ است. کار دیگری که در این مرحله انجام می‌شود ترکیب و تشخیص اعداد است. در این مرحله چند نوع برچسب جدید تعریف می‌شوند که با ترکیب علامات (':' و '/' و '.' و '-') با اعداد ترکیب جدید با برچسب جدید ایجاد می‌کند. همچنین اعداد فارسی را که بینشان 'و' می‌باشد را با هم ترکیب کرده و یک عدد مرکب ایجاد می‌کند.

در جدول (۱) برچسب‌های بکار رفته در این مرحله آورده شده است.

جدول ۱: برچسب های مقوله های نحوی اضافه

کد مقوله نحوی	توضیح	کد مقوله نحوی	توضیح
FNo	عدد اعشاری	timeNo	ترکیب زمان (۸:۵۶)
moneyNo	عدد پول (۱۲۳.۲۳۴.۰۰۰)	dateNo	ترکیب تاریخ (۱۹۹۹/Jun/۱۲)

۵-۱-۲-۲- برچسب زنی اسامی خاص

برای ارتقاء و افزایش کارایی برچسب‌گذاری‌ها از پایگاه داده دیگری استفاده می‌شود که در بر دارنده ماه‌ها، روزها، عناوین اشخاص، اماکن، واحدهای پولی و اسامی خاص‌اند. این پایگاه داده کلمات ترکیبی(مثل «آلیل پیروپیل دی سولفید») را نیز در بر دارد که در صورت وجود چنین کلمه‌ای در متن، آن جایگزین چندین کلمه برچسب‌گذاری شده قبلی با توجه به مکانش می‌شود. به این ترتیب به با اضافه شدن این برچسب‌های جدید، برچسب‌گذاری ارتقاء می‌یابد.

۵-۱-۲-۳- ریخت شناسی[۱۳] اسامی خاص

در زبان فارسی بعضی از اسامی خاص از ساختار خاص خود پیروی می‌کنند مثلاً نام‌های افراد و یا نواحی، ممکن است از وندهای خاصی نظیر آنچه که در «یعقوب + زاده»، «پور + حسن»، «پونه سادات مرتضوی» یا «علی آباد» مشاهده می‌شود پیروی کنند. علاوه بر موارد ذکر شده، یک اسم خاص هم می‌تواند خود، ترکیبی از چند اسم خاص دیگر باشد.

این سیستم برای تشخیص اسامی خاص از روی ساختار آنها، از پایگاه داده‌ای از وندها بهره می‌برد. این پایگاه داده علاوه بر اینکه نوع هر وند را نشان می‌دهد محدودیت وند را هم در پذیرش نوع ریشه‌ی کلمه(مثلاً اسم خاص فقط می‌تواند باشد و یا ...) و نیز نقش سازنده وند را هم نشان می‌دهد.

پس از مرحله ثانویه برچسب گذاری، به بررسی ساختار کلمات به دو شکل مجزا و داخل متن پرداخته می‌شود. در مرحله‌ی اول هر کلمه از لحاظ اجزای تشکیل دهنده‌اش و ترکیب حروفی خاص که در ابتدا و انتهای آن آمده است بررسی می‌شود. مثلاً کلمه «گلشیفته» که در مجموع کلمات ذخیره شده وجود ندارد و عبارتست از «گل + شیفته»، گل هم که به عنوان پیشوند چسبیده اسم ساز شناخته شده است و برچسب شیفته هم جز برچسب‌های مجاز ترکیب با گل است پس اسم خاص تشخیص داده می‌شود.

در مرحله دوم، کلمه‌ی قبل و بعد کلمه مذکور -که اینک ممکن است به دلیل مرحله قبل، نوع آن تغییر کرده باشد- را نگاه می‌کند. در فارسی گاهی اعضای یک کلمه به صورت کلمات مجزا با فاصله نوشته می‌شوند. بنابراین اسم خاصی چون «معین زاده» به صورت مجزا خوانده و چه بسا نادرست برچسب زده می‌شود. در این مرحله اینگونه خطاها تصحیح می‌شوند.

پس از انجام این پیش پردازش‌ها وارد مرحله پردازش اصلی می‌شویم که در ادامه به آن پرداخته می‌شود.

۵-۲- شناسایی موجودیت و برچسب معنایی آن

پس از مرحله پیش پردازش نوبت به شناسایی موجودیت-های نامدار با استفاده از متن برچسب گذاری شده می‌رسد. در این مرحله ابتدا NEها با استفاده از روش‌های بکار برده شده که در ادامه مطرح شده‌اند حدس زده می‌شوند، سپس صحت حدس‌های زده شده بررسی می‌شود و نهایتاً NEها و نوعشان مشخص می‌شود.

۵-۲-۱- تشخیص مبتنی بر الگو

برای تشخیص یک کلمه در متنش، می‌توان از الگو استفاده کرد. در این سیستم نیز توجه به الگویی که هر کلمه در آن قرار گرفته به عنوان مبنای تشخیص بکار می‌رود. الگوها به صورت مجموعه‌ای از کلمات و برچسب‌هایی هستند که در متن بدنبال هم می‌آیند. برای هر نوع از موجودیت‌های نامدار که در ساختار سلسله مراتبی بخش قبل آورده شده است می‌توان یک یا چندین الگو در نظر گرفت.

الگوها به طور عام به صورت «دسته ×» و یا «دسته ××» ظاهر می شوند. × در الگوی اول یک موجودیت یک کلمه ای و

×× در الگوی دوم جایگاه یک موجودیت چند کلمه ایست. مثلاً «دولت ×» برای نام کشور و «نهاد ××» برای نام یک نهاد بکار می‌رود. در اینصورت عباراتی مثل "...دولت ایران.." و یا "...نهاد نمایندگی مقام معظم رهبری در دانشگاه ها..." با این الگوها مطابقت خواهند کرد.

مجموعاً در این سیستم، جستجو برای تشخیص موجودیت نامدار با کمک الگوها به این صورت انجام می‌گیرد که برای تک-تک عناصر برچسب گذاری شده در متن (با در نظر داشتن ضمایر چسبیده به کلمات و نیز ادات جمع)، هم برای خود کلمه و هم برای برچسب آن بررسی می‌شود که آیا در هیچ یک از الگو-ها صدق می‌کنند یا نه و به این ترتیب مجموعه‌ای از اسامی خاص و نوع آنها (که عنوان مقوله هر الگو است) مشخص می-شود.

۵-۲-۲-تأثیر کلمات جمع در اسامی خاص

گاهی مجموعه‌ای از چندین موجودیت نامدار متعلق به یک مقوله بدنبال یکدیگر ذکر می‌شوند. مثلاً داریم «شرکت‌های ایزایران و سافت نت با همکاری هم نرم افزار جدید ایجاد نمودند.». در اینجا هر یک از مجموعه کلماتی که زیرشان خط کشیده شده است به عنوان یک اسم خاص برای شرکت بکار می‌رود.

در دو حالت مجموعه‌ای از موجودیت‌های خاص بدنبال یکدیگر را سیستم در یک دسته قرار می‌دهد. حالت اول زمانی است که کلمه‌ای که به عنوان شناساینده موجودیت، در همسایگی آن آمده است به صورت جمع آورده شده باشد و حالت دیگر وجود صفت شمارشی پیش از آن، مثلاً «دو شرکت ایزایران و سافت نت». در حالت دوم سیستم بدنبال اعداد شمارشی استفاده نشده در الگوها پیش از شناساینده‌ها می‌گردد تا آنها را به صورت جمع دیده و بجای یک اسم نامدار همچون شناساینده‌های جمع، بدنبال مجموعه‌ای از اسامی نامدار بگردد.

۵-۲-۳- نا کافی بودن الگو

این گونه استخراج موجودیت نامدار سه مشکل دارد، اول اینکه چه تعداد از دنباله کلمات، موجودیت نامدار درست را تشکیل می‌دهند یا به عبارتی مرز موجودیت نامدار کجاست. مثلاً برای نام شرکت که در الگو به صورت ** مشخص شده است چه دنباله‌ای نام شرکت است.

مشکل دوم این است که آیا کلمه تشخیص داده شده همان موجودیت نامدار است یا موجودیت ناخواسته[۱۴] می‌باشد. مثلاً برای «تاریخ هجری قمری» با استفاده از الگو «تاریخ ×»، هجری را به عنوان تاریخ در نظر می‌گیرد در حالی که تاریخ فرمت خاص خود را دارد که با برچسب محاسبه شده dateNo یا ترکیبی از برچسب‌های مشخص حاصل می‌شود.

مشکل سوم، ناشناس ماندن برخی از موجودیت‌هایی است که در هیچ الگویی صدق نکرده‌اند. در ادامه روش‌هایی برای ابهام زدایی و رفع این مشکلات احتمالی آورده شده است.

۵-۲-۴- ابهام زدایی

۵-۲-۴-۱- ابهام زدایی با توجه به تکرار رخداد

برای اولین مشکل مطرح شده یعنی تعیین مرز اسم نامدار، کار ابتدایی که انجام شد بررسی بلندترین دنباله اسمی تکرار شده در متن است. چرا که اسامی خاص به احتمال زیاد جای دیگری از متن باز تکرار می‌شوند. در این صورت، اگر هر دو متعلق به یک مقوله بودند، احتمال صحت افزایش می‌یابد. از آنجایی که یک موجودیت ممکن است در جای دیگر به صورت خلاصه شده‌ای تکرار شود پس در صورت ترکیبی بودن، ترکیب-های مختلف آن با حفظ ترتیب، برای تکرار آزمایش می‌شود.

۵-۲-۴-۲- ابهام زدایی با ترکیب کلمات

برای رفع دومین مشکل مطرح شده که در بهبود مشکل اول هم دخیل است، با بهره‌وری از پایگاه داده‌ای از ترکیب‌های برچسبی و کلمه‌ای ممکن برای هر مقوله به عنوان مکملی برای الگوها، مشکل تقلیل می‌یابد. به عبارتی هر کلمه پیش از اینکه جزوی از یک موجودیت نامدار محسوب شود امکان موجودیت بودن آن بررسی می‌شود. پیش از این در مورد پایگاه داده طراحی شده برای ترکیب کلمات توضیح داده شد. در جایی که سیستم، بدنبال موجودیت نامدار مناسب برای جایگزینی در الگو می‌گردد، آن موجودیتی را می‌پذیرد که خود کلمه و یا برچسب نحوی که الگو برای آن مطرح است، برای موجودیت صدق کند. مثلاً برای «مؤسسه آموزشی دانشجو و ایران» با توجه به الگوی «+مؤسسه +آموزشی ××» برای مقوله سازمان←آموزشی تک تک کلمات پس از آموزشی را تا جایی که خود کلمه و یا برچسب آن برای مقوله آموزشی صدق می‌کنند می‌پذیرد.

۵-۳- پالایش و تولید خروجی نهایی

پس از کشف موجودیت‌ها و اعمال برچسب معنایی به آنها، سیستم وارد مرحله پردازش نهایی می‌شود. در این مرحله با توجه به ساختار سلسله مراتبی سیستم، موجودیت‌های نامدار تکرار شده در یک مقوله و یا در مقوله‌های سطح پایین‌تر (گره

فرزند) کشف شده و حذف می‌شوند. همچنین موجودیت‌هایی که درون موجودیتی دیگر قرار گرفته‌اند با توجه به مکان آنها در متن شناسایی و حذف می‌گردند. مثلاً «انجمن شرکت‌های بازرگانان نمونه»، ممکن است «بازرگانان نمونه» را از مقوله شرکت ببیند در حالی که «شرکت‌های بازرگانان نمونه» از مقوله انجمن، موجودیت خاص مورد نظر است. پس سیستم با مشاهده چنین شرایطی موجودیت نامدار تکرار شده در موجودیت نامدار دیگر را حذف می‌کند. نهایتاً مجموعه موجودیت‌های شناسایی شده در مقوله مناسبشان دسته بندی شده و خروجی با برچسب‌های معنایی تولید می‌شود.

۶- آزمون و ارزیابی

سیستم به ازای متون مختلف و در بر دارنده تقریباً تمامی برچسب‌های معنایی ممکن آزمایش شد. پس از مقایسه‌ی نتایج حاصله از برچسب‌گذاری سیستم بر داده‌های تستی با برچسب-های معنایی تولید شده به صورت دستی، دقت[۱۵] سیستم ۷۲٪ و یادآوری[۱۶] آن ۷۶٪ گزارش شد.

علت عمده عدم تشخیص، به تشخیص ندادن مرز درست برخی از موجودیت‌های نامدار در متن باز می‌گشت. به عنوان نمونه، رفتار سیستم در مقابل عبارت «شرکت فرارسانه اندیشه ساز» تشخیص «فرارسانه» به عنوان موجودیت نامداری از «شرکت» بود. علت رخداد این خطا، عدم تکرار عبارت در جای دیگری از متن و نیز نامرتبط بودن برچسب‌های نحوی واژگان تشکیل دهنده عبارت بود. سایر رخدادهای عدم تشخیص موجودیت نامدار، بدلیل صدق نکردن آن در هیچ گونه از الگوها می‌باشد.

اشتباه در تشخیص موجودیت‌های نامدار هم اکثراً یا به میزان تشابه‌شان با ظاهر موجودیت‌های نامدار بازمی‌گشت که ریخت شناسی منجر به این خطا می‌شد و یا به استفاده از موجودیت در جایی که انتظار مشاهده موجودیت نامدار می‌رفت. به عبارتی سیستم با ابهام در تشخیص مواجه می‌شد. گاهی هم خطا در برچسب زن نحوی، منجر به انتشار خطا در نتیجه نهایی سیستم می‌شد.

۷- نتیجه گیری

در اینجا سیستمی برای تشخیص و دسته بندی موجودیت-های نامدار در زبان فارسی معرفی شد که برای رسیدن به مقصود خود از برچسب‌گذاری نحوی و ویژگی‌های درونی و متنی واژگان بهره می‌برد. این سیستم به طور کامل پیاده سازی و آزمایش شده است و قابل استفاده به عنوان مکملی برای کاربردهایی از قبیل برچسب‌گذاری‌های نحوی، استخراج اطلاعات و درک مطلب و خلاصه سازی ماشین می‌باشد.

نتایج حاصله از آزمایش سیستم بجز شرایط بروز ابهام در متن خوب گذارش شده است. برای رفع مشکل مرز موجودیت نامدار که از روش‌های رفع ابهام مطرح شده در متن رفع نمی-شود، سیستم به عنوان مکمل، نیاز به قطعه‌ساز[۱۷]ی دارد که اجزای جمله را از یکدیگر تفکیک نماید. علاوه بر آن استفاده از الگوریتم‌های یادگیری ماشین که از متون حاشیه نویسی استفاده می‌کنند می‌تواند در افزایش دقت سیستم تأثیر چشم-گیری داشته باشد.

مراجع

[۱] موسوی، نفیسه سادات؛ ثانی، غلامرضا قاسم؛ "بکارگیری دسته بندی کننده و رتبه بندی کننده آنتروپی بیشینه در فرایند تعیین"، کنفرانس ملی انجمن کامپیوتر ایران، دوره ۱۴، تهران، ۱۳۸۷.

[2] Grishman, R., Sundheim, B. *Message Understanding Conference - 6: A brief history*, COLINGS, 1996.

[3] Zhang, Li, Pan, Yue, Zhang, Tong. *Focused Named Entity Recognition using Machine Learning*. ACM Press, pp. 281-288, 2004.

[4] Cohen, William W., Sarawagi, Sunita. *Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods*. Conference on Knowledge Discovery in Data, pp. 89-98, 2004.

[5] Nadeau, David, Sekine, Satoshi. "A survey of named entity recognition and classification.", National Research Council Canada/New York University, 2006.

[6] Negri, Matteo, Magnini, Bernardo. *Using WordNet Predicates for Multilingual Named Entity Recognition*, The Second Global Wordnet Conference, pp. 169-174,2004.

[7] mikheev, Andrei, Moens, Marc, Grover, Claire. *Named Entity Recognition without Gazetteers*, Proceedings of EACL, Vol. 19, Bergen, Norway, pp. 1-8, 1999.

[8] Cucerzan, Silviu, Yarowsky, David. *Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence*, Proceedings of EMNLP, pp. 90-99, 1999.

[9] Resnick, Ed., P. *Internet Message Format (rfc5322)*, Network Working Group, 2008.

[10] Berners-Lee, T., Masinter, L., McCahill, M. *RFC1738 - Uniform Resource Locators (URL)*. Network Working Group, December 1994, http://www.faqs.org/rfcs/rfc1738.html

[11] Uniform Resource Identifer (URI) Schemes, *IANA*, ICANN, ptc/10-11-2009, http://www.iana.org/assignments/uri-schemes.html.

[12] Shamsfard, Mehrnoush, Fadaee, Hakimeh. *A Hybrid Morphology-Based POS Tagger for Persian*, European Language Resources Association (ELRA), Marrakech, Morocco, 2008.

[۱] Named Entity Recognition(NER)

[۲] Supervised Learning

[۳] Semi-supervised Learning

[۴] Unsupervised Learning

[۵] **پایگاه داده لغوی انگلیسی است که در آن اسم‌ها، فعل‌ها، صفات و قیود به مجموعه‌ای از مترادف‌های ادراکی**(synsets) **گروه‌بندی می‌شوند که هر کدام بیانگر مفهوم مجزایی‌اند.**

[۶] Tokenization

[۷] Local-Part

[۸] Internet Domain

[۹] Scheme

[۱۰] Internet Assigned Numbers Authority

[۱۱] POS Tagger

[۱۲] Tagset

[۱۳] Morphology

[۱۴] noise

[۱۵] Precision

[۱۶] Recall

[۱۷] Chunker