
بسم الله الرحمن الرحیم

تعهدنامه

اینجانب **طوبی فدائی تبریزی** دانشجوی دوره **کارشناسی ارشد** رشته مهندسی کامپیوتر دانشکده مهندسی دانشگاه فردوسی مشهد نویسنده پایان‌نامه **استخراج حقایق از متون فارسی درقالب RDF** تحت راهنمایی **پروفسور محسن کاهانی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود و یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه فردوسی مشهد می باشد و مقالات مستخرج با نام "دانشگاه فردوسی مشهد" و یا "Ferdowsi University of Mashhad" به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند در مقالات مستخرج از رساله رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است، اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه فردوسی مشهد می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

آزمایشگاه ناوری وب

پایان‌نامه کارشناسی ارشد

استخراج حقایق از متون فارسی در قالب RDF

تهیه و تنظیم: طوبی فدائی تبریزی

استاد راهنما: پروفسور محسن کاهانی

بهمن ماه ۱۳۹۳

چکیده

با توجه به حجم عظیم دانش و اطلاعات بشر و رشد روزافزون مستندات در زمینه‌های مختلف، پردازش زبان‌های طبیعی و تبدیل متون به دانش قابل فهم برای ماشین، مورد توجه قرار گرفته است. با استفاده از سیستم‌های استخراج اطلاعات می‌توان بطور خودکار پایگاه دانشی ساخت‌یافته از متون ایجاد کرد. در واقع هدف یک سیستم استخراج اطلاعات، استخراج حقایق از متون غیرساخت‌یافته و نمایش آن‌ها در قالب‌های ساخت‌یافته مانند سه‌گانه‌های RDF می‌باشد. اگر حقایق در قالب معنایی RDF نگاشت شوند، می‌توان اطلاعات مورد نیاز را با ساخت و ارسال پرس‌وجوهای SPARQL روی پایگاه دانش بدست آورد. در این پایان‌نامه، روشی برای استخراج آزاد حقایق از متون زبان فارسی پیشنهاد شده است که در آن استخراج حقایق در سطح جمله و بر اساس تشخیص افعال و روابط وابستگی بین اجزای جمله انجام می‌شود.

راه‌کار پیشنهادی، حقایق اصلی را بر اساس فعل و حقایق فرعی را بر اساس روابط بین گروه‌های اسمی جمله استخراج و برای تبدیل به قالب RDF آماده‌سازی می‌کند. برای نگاشت حقایق در قالب معنایی RDF، URI قسمت‌های نهاد، مسند و گزاره یک حقیقت با استفاده از شبکه واژگان و ویکی‌پدیا شناسایی می‌شود. در نتیجه در راه‌کار پیشنهادی شبکه واژگان فردوس‌نت بصورت خودکار بر اساس شبکه واژگان انگلیسی ایجاد می‌شود. نتایج حاصل از ارزیابی نشان می‌دهد که روش پیشنهادی در استخراج حقایق موفق بوده و باعث بهبود دقت و فراخوانی نسبت به سیستم‌های موجود می‌شود. علاوه براین سیستم پیشنهادی حقایق را در قالب معنایی RDF استخراج می‌کند.

کلمات کلیدی: پردازش زبان طبیعی، استخراج آزاد اطلاعات، استخراج دانش، وب معنایی، RDF

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل اول:

مقدمه

فصل ۱- مقدمه

در این فصل ابتدا به تعریف مساله‌ی مورد توجه این تحقیق پرداخته می‌شود. در ادامه راه‌حل پیشنهادی و نوآوری‌های پایان‌نامه برای حل مسئله‌ی مورد نظر بیان خواهد شد. در پایان فصل نیز، ساختار پایان‌نامه ارائه خواهد شد.

۱-۱- تعریف مساله

اسناد متنی دارای اطلاعات با ارزشی از قبیل اخبار و رویدادهای مختلف، مقالات علمی، گزارش‌های اداری و مالی و .. می‌باشند و می‌توانند منابع غنی برای تصمیم‌گیری باشند. چنانچه این اطلاعات، به شکل ساخت‌یافته در یک پایگاه دانش ذخیره شده باشند، می‌توانند به صورت ساده‌تری مدیریت شوند، و مورد پرس‌وجو قرار گیرند. اگر پایگاه دانشی با قابلیت اطمینان و دقت بالا تهیه شود، بسیاری از مشکلات بازیابی، پردازش و تحلیل اطلاعات رفع می‌گردد. برای ایجاد چنین پایگاه دانشی از سیستم‌های استخراج اطلاعات استفاده می‌شود.

هسته اصلی سیستم‌های استخراج اطلاعات، کاوش کلمات کلیدی و اصلی متن می‌باشد. کاوشگر اطلاعات با استفاده از پردازش متن از کلمات نامربوط چشم‌پوشی می‌کند و کلمات مناسب و حاوی مفهوم و همچنین روابط بین آنها را کشف می‌کند [1, 2]. استخراج اطلاعات در سه سطح موجودیت، جمله و صفحه انجام می‌شود [3]:

- سطح موجودیت: استخراج اطلاعات در سطح موجودیت، ساده‌ترین مدل استخراج با فرض استقلال موجودیت‌های متون می‌باشد. در این سطح رابطه‌ای میان دو موجودیت مستقل از موجودیت‌های دیگر و همچنین مستقل از تشخیص کلمه کلیدی رابطه مشخص می‌شود. همان طور که اشاره شد، در این سطح موجودیت‌ها کاملا مستقل در نظر گرفته می‌شوند،

لذا برای کشف روابط بین دو موجودیت، باید اطلاعات کافی در مورد آنها وجود داشته باشد.

- سطح جمله: سیستم‌های استخراج اطلاعات در سطح جمله، برای توصیف یک معنی خاص، کلمات را در سطح جمله مستقل از یکدیگر نمی‌دانند. در این سطح رابطه جفت موجودیت‌ها در سطح یک جمله بررسی می‌شوند. نوع رابطه با توجه به کلمات‌های اطراف موجودیت‌ها، مشخص می‌شود. با استفاده از روابط نحوی بین کلمات جمله، کلمه کلیدی روابط مشخص می‌شود. بیشتر سیستم‌ها، استخراج اطلاعات را در سطح جمله انجام می‌دهند.
- سطح صفحه: ممکن است یک جمله در سطح یک صفحه وب و یا یک سند متنی، کاملاً مستقل نباشد و به رابطه جملات، در سطح صفحه نیاز باشد. در این حالت با توجه به استنتاج‌های پیوندی و استدلال‌های قوی، می‌توان روابط را در سطح صفحه استخراج نمود.

شناسایی روابط در استخراج اطلاعات می‌تواند به دو روش هدفمند مبتنی بر خروجی (بسته)[۱] و عمومی مبتنی بر ورودی (آزاد)[۲]انجام شود [2]. در روش هدفمند مبتنی بر خروجی با توجه به یک مجموعه روابط مشخص، نمونه‌ها جمع‌آوری می‌شوند. این سیستم‌ها مستقل از منبع ورودی هستند. در این روش، فقط حقایق مربوط به روابط خاص کشف می‌شود. این روابط باید از قبل مشخص شوند و برای هر نوع رابطه، به مجموعه داده آموزشی نیاز است. در نتیجه افزودن یک رابطه جدید هزینه بالایی دارد. در روش عمومی مبتنی بر ورودی، تمرکز روی منبع یا دامنه ورودی است و به همه روابط ممکن، رسیدگی می‌شود. در این روش تحلیل نحوی عمیق‌تری بر روی متن انجام می‌شود [1, 4, 5].

---

[۱] Output-oriented targeted(“closed”)

[۲] Input-oriented generic (“open”)

خروجی روش استخراج آزاد، حقایقی می‌باشند که به صورت سه‌تایی بیان شده‌اند. در این سیستم‌ها، حقایق برای نگاشت به مدل داده‌ای RDF[۳] تهیه نشده‌اند. برای مثال ممکن است یک جمله یا ترکیبی از کلمات و حروف اضافه در یکی از آرگومان‌های حقیقت وجود داشته باشد [6, 7, 8]. اگر این حقایق به قالب معنایی RDF تبدیل شوند،با استفاده از لینک به داده‌های پیوندی[۴] و سایر هستان‌شناسی‌ها[۵] و توضیحات معنایی می‌توان اطلاعات بیشتری در مورد آرگومان‌ها و رابطه حقایق بدست آورد. با استفاده از این اطلاعات می‌توان استنتاج‌های لازم را انجام داد و حقایق جدید و مفیدی به پایگاه دانش اضافه شود و مورد تحلیل و بررسی قرار گیرد.

سیستم‌های استخراج اطلاعات، حقایق را از طریق پنج روش، روش مبتنی بر قاعده[۶] [1, 9, 10, 11]، روش مبتنی بر الگو[۷] [1, 7, 12, 13, 14]، روش مبتنی بر استدلال [2]، روش مبتنی بر پردازش زبان طبیعی[۸] [15, 16, 17] و روش مبتنی بر هستان‌شناسی [18, 19, 20, 21, 22]، شناسایی می‌کنند.

روش مبتنی بر قاعده، امکان ایجاد و استنتاج خودکار برای استخراج حقیقت از جداول، لیست‌ها، فیلدهای فرم و عناصر دیگر نیمه ساخت‌یافته فراهم می‌کند. در روش مبتنی بر الگو، حقایق با اعمال الگوهای متنی بر روی متون غیرساخت‌یافته استخراج می‌شوند. روش مبتنی بر استدلال، با انجام عمل اضافی استدلال بر روی مجموعه فرضیات حقایق جمع‌آوری شده،حقایق جدیدی به پایگاه دانش اضافه می‌کند. روش مبتنی بر پردازش طبیعی، با توجه به پیچیدگی ساختار متن از NLP

---

[۳] Resource Description Framework

[۴] Linked Data

[۵] Ontology

[۶] Rule-based Method

[۷] Pattern-based Method

[۸] Natural Language Processing

سطحی وNLP عمقی[۹]استفاده می‌کند. روش مبتنی بر هستان‌شناسی OBIE[۱۰]، یک روند شناسایی مفاهیم، خصیصه‌ها و روابط بیان شده در هستان‌شناسی، از متن و منابع دیگر است. در این روش موجودیت‌ها می‌توانند با بررسی اسناد، توضیحات خود را در طی روند استخراج اطلاعات غنی‌تر کنند.

یکی از چالش‌های استخراج اطلاعات از متون زبان فارسی، نبود مجموعه داده آموزشی برای تولید خودکار الگوهای متنی است و از طرفی تولید الگوها و یا قواعد به صورت دستی نیز هزینه‌بر می‌باشد. یکی دیگر از چالش‌های موجود برای زبان فارسی، نبود هستان‌شناسی کاملی است که بتوان از آن در استخراج حقایق استفاده کرد. از طرفی ابزارهای برچسب‌زنی معنایی با دقت قابل قبول وجود ندارد. مسئله‌ی ابهام علاوه بر کلمات هم‌معنی در مورد افعال مرکب هم می‌باشد و همچنین پیش فرض خاصی در مورد قالب جملات وجود ندارد.

۱-۲- راه‌حل پیشنهادی

هدف اصلی این تحقیق، استخراج آزاد حقایق از متون زبان فارسی برای تبدیل به قالب معنایی RDF با دقت و فراخوانی مناسب است. با توجه به روش‌ها و سیستم‌های استخراج اطلاعات و چالش‌های آنها که در بخش مرور ادبیات مطرح خواهد شد، برای تحقق این هدف، از روش استخراج آزاد اطلاعات در سطح جمله و تجزیه وابستگی برای پردازش زبان طبیعی استفاده شده است.

اسنادی که برای استخراج حقایق به سیستم داده می‌شوند، باید حاوی اطلاعات واقعی باشند و اطلاعات نادرست در آنها وجود نداشته باشد. علاوه بر این باید اسناد بصورت رسمی بیان شده باشند و جملات آنها از نظر گرامری صحیح باشند. به عبارت دیگر نباید اسناد ورودی دارای اصطلاحات و جملات محاوره‌ای باشند.

---

[۹] Deep NLP (DNLP)

[۱۰] Ontology-based Information Extraction

در بخش پیش‌پردازش، اسناد و صفحات وب دریافت می‌شوند. سپس متن به پاراگراف‌ها، جملات، کلمات و عبارات تقسیم می‌شود. برچسب گذاری اجزای کلام[۱۱] POS و انتساب مقوله نحوی به هر کلمه در پیکره متنی، از جمله کارهای دیگری است که در این مرحله انجام می‌شود. با توجه به چالش‌های مربوط به استخراج حقایق از متون زبان فارسی که در تعرف مساله، به آنها اشاره شد، از تجزیه وابستگی برای یافتن روابط وابستگی بین عناصر جمله استفاده می‌شود. مرجع ضمایر و کلمات مانند اختصارات در پیش پردازش کشف می‌شود.

راه‌کار پیشنهادی بر اساس تجزیه وابستگی می‌باشد. در مرحله اول استخراج حقایق، جمله (ساده یا مرکب) از دو منظر بررسی می‌شود و حقایق آن با استفاده از یک سری الگوی نحوی بر اساس گرامر زبان فارسی استخراج می‌شود.

- حقایق اصلی جمله که بر اساس فعل جمله شناسایی می‌شوند.
- حقایق فرعی که بر اساس ارتباطات بین گروه‌های اسمی کشف می‌شوند.

سیستم پیشنهادی با استفاده از ارتباط بین کلمات جمله، حقایق را استخراج می‌کند. برای کشف حقایق اصلی، کلماتی که با فعل ارتباط دارند، مورد بررسی قرار می‌گیرند. فاعل جمله بعنوان نهاد حقیقت و فعل به عنوان مسند حقیقت انتخاب می‌شود. گزاره هر حقیقت با توجه به نوع وابستگی نحوی کلمات مشخص می‌شود. اگر برای حقیقتی، گزاره شناسایی نشود، گزاره آن تهی است.

برای کشف حقایق فرعی، روابط بین گروه‌های اسمی بررسی می‌شود. گروه‌های اسمی شامل اطلاعات مربوط به بدل، ترکیبات مضاف و مضاف‌الیه و موصوف و صفت می‌باشند. حقایق با استفاده از الگوهای نحوی مربوط به گروه‌های اسمی، شناسایی می‌شوند.

---

[۱۱] Part of Speech

سیستم پیشنهادی حقایق را برای نگاشت به قالب معنایی RDF آماده سازی می‌کند. در نتیجه تا حد امکان از کلمات و یا حروف اضافی خودداری می‌شود؛ همچنین قسمت‌های نهاد، مسند و گزاره حقایق، شکسته و کوچک شوند. بنابراین حروف ربطی در قسمت‌های مختلف یک حقیقت مورد بررسی قرار می‌گیرند. حقایق جدید با شکسته شدن قسمت‌های نهاد، مسند و گزاره حقایق بدست می‌آید. در این بخش گروه‌های حرف اضافه بررسی می‌شوند و با توجه به نقش نحوی آنها ابعاد مکان و زمان حقایق کشف می‌شود.

در راه‌کار پیشنهادی برای شناسایی موجودیت‌ها و روابط به یک شبکه واژگان، نیاز است. در این راستا با استفاده از شبکه واژگان انگلیسی [23]، شبکه واژگان فارس‌نت [24] و فرهنگ لغات مختلف و فرهنگ واژگان مترادف و متضاد، شبکه واژگان فردوس‌نت به صورت خودکار ایجاد شده است.

در انتها سیستم پیشنهادی با سیستم‌های موجود مقایسه شده است. نتایج حاصل از ارزیابی کیفی و کمی نشان می‌دهد که روش پیشنهادی باعث بهبود دقت و فراخوانی نسبت به سیستم‌های موجود می‌شود. همچنین سیستم پیشنهادی، حقایق را در قالب RDF استخراج می‌کند، تا بتوان از توضیحات معنایی اجزای حقیقت برای استنتاج و تحلیل استفاده نمود .

۱-۳- نوآوری‌های راه‌حل پیشنهادی

نوآوری سیستم پیشنهادی در موارد زیر خلاصه می‌شود:

۱. طراحی یک روش استخراج آزاد حقایق از متون زبان فارسی با بکارگیری از تجزیه وابستگی
۲. استخراج حقایق در قالب معنایی RDF
۳. ایجاد خودکار شبکه واژگان فردوس‌نت

۱-۴- ساختار پایان‌نامه

این پایان‌نامه دارای پنج فصل می‌باشد. در فصل نخست، کلیات، اهداف و ضرورت انجام تحقیق ارائه شد. در فصل دوم، در ابتدا مراحل استخراج اطلاعات و روش‌های استخراج اطلاعات شرح داده می‌شود و سپس برخی سیستم‌های استخراج آزاد اطلاعات بررسی می‌شوند. در انتهای این فصل به کارهای انجام شده در استخراج اطلاعات از متون فارسی و ایجاد شبکه واژگان پرداخته می‌شود.

در فصل سوم، جزئیات راه حل پیشنهادی برای استخراج حقایق از متون زبان فارسی و آماده‌سازی حقایق برای تبدیل به قالب معنایی RDF توضیح داده می‌شود. همچنین چگونگی ایجاد خودکار شبکه واژگان براساس شبکه واژگان انگلیسی شرح داده می‌شود. در ادامه‌ی فصل به چگونگی تولید صفحات RDF از شبکه واژگان فردوس‌نت پرداخته می‌شود.

در فصل چهارم، جزئیات پیاده‌سازی روش پیشنهادی به همراه نحوه‌ی ارزیابی و نتایج حاصل از آن بیان شده است. در پایان، نتیجه‌گیری و کارهای آتی در فصل پنجم ارائه می‌شود.

فصل دوم:

مرور ادبیات

فصل ۲- مرور ادبیات

۲-۱- مقدمه

سیستم پیشنهادی این پایان‌نامه، برای استخراج اطلاعات از متون زبان فارسی استفاده می‌شود. هدف از ارائه این سیستم حل برخی چالش‌هایی است که سیستم‌های استخراج اطلاعات بخصوص در زبان فارسی با آن روبرو هستند. بنابراین در این فصل به بررسی سیستم‌های استخراج اطلاعات و چالش‌های آنها پرداخته می‌شود. در بخش اول روش‌های استخراج اطلاعات توضیح داده می‌شود. در بخش دوم سیستم‌های استخراج آزاد اطلاعات بررسی و مقایسه می‌شوند. در بخش سوم این فصل، استخراج اطلاعات در متون فارسی مورد بحث و بررسی قرار می‌گیرد. در انتها به کارهای انجام شده در ایجاد شبکه واژگان پرداخته می‌شود.

۲-۲- مفاهیم پایه

در این بخش به مفاهیم پیش‌نیاز در استخراج اطلاعات پرداخته می‌شود.

۲-۲-۱- استخراج اطلاعات

اسناد متنی دارای اطلاعات با ارزشی هستند، اما با توجه به میزان قابل توجه متون که هر روز به حجم الکترونیکی آنها افزوده می‌شود، پیدا کردن اطلاعات مربوط و لازم از منابع مختلف برای کاربر مشکل است، که آنها را بخواند، بفهمد و با هم ادغام کند. اگر همه این اطلاعات، به شکل ساخت‌یافته‌تری در یک پایگاه دانش ذخیره شده باشند، می‌توانند به صورت ساده‌تری مدیریت شوند، و مورد پرس‌وجو قرار گیرند. استخراج اطلاعات به منظور تبدیل متن به اطلاعات قابل استفاده از منظر

ماشین استفاده می‌شود. سیستم‌های استخراج اطلاعات با کمک نرم افزارها و روش‌های پردازش متون طبیعی، متون را برای استخراج اطلاعات مورد نیاز، پردازش می‌کنند و می‌توانند اطلاعات بدست آمده را در پایگاه دانش ذخیره کنند [25, 26].

برخی از کاربردهای مهم سیستم‌های استخراج اطلاعات عبارتند از [27, 2]:

- ایجاد یک پایگاه دانش یا دایره‌المعارف معنایی با دقت بالا و قابل خواندن برای ماشین جهت استفاده در استنتاج‌های سیستم‌های پرسش‌و پاسخ
- ابهام‌زدایی از موجودیت‌های عبارات متنی بوسیله نگاشت به موجودیت‌های نام‌دار موجود در پایگاه دانش
- جستجوی معنایی بر مبنای موجودیت- رابطه بر روی حجم انبوه مستندات یا صفحات وب
- استفاده در سیستم‌های پرسش و پاسخ و پاسخ‌گویی به سوالات زبان طبیعی با توجه به موجودیت‌ها و روابط آنها در سوالاتی مانند کی، کجا، چه موقع و ...
- سیستم‌های ترجمه ماشینی، سازماندهی، نگهداری و رشد خودکار پایگاه دانش، تشابه و ارتباط معنایی بین متون، نظرکاوی و سایر عملیات متن‌کاوی

چالش‌های اساسی در سیستم‌های استخراج اطلاعات مربوط به ابهام، مقیاس‌پذیری و ارزیابی می‌باشد.

- ابهام: موجودیت‌ها و روابط، دارای نام‌های مختلفی می‌باشند؛ تشخیص نام‌های هم‌معنی موجودیت‌ها و روابط بین آنها، به عنوان ابهام‌زدایی از اطلاعات استخراج شده، مطرح می‌گردد [26]. برای نمونه، به طور متوسط ۲,۹ نام برای هر موجودیت، در بین ۸۰ موجودیت پرکاربرد و ۴,۹ نام برای هر رابطه، در بین ۱۰۰ رابطه پرکاربرد در یک مجموعه داده استفاده می‌شود [28].

- مقیاس‌پذیری: استخراج اطلاعات و برنامه‌های کاربردی متن کاوی با مقدار بسیار زیادی از اطلاعات متنی با ارزش روبرو هستند، به عبارت دیگر باید بتوانند اطلاعات را از میلیون‌ها، و در بعضی موارد میلیاردها سند استخراج کنند. متاسفانه استخراج موجودیت‌ها و روابط از اسناد از نظر محاسباتی یک کار سنگین و گران می‌باشد. حتی یک استخراج اطلاعات ساده ممکن است، روزها یا هفته‌ها زمان برای پردازش مجموعه بزرگ، نیاز داشته باشد [30 ,29].
- ارزیابی: ارزیابی اطلاعات استخراج شده را می‌توان از دو جنبه بررسی نمود؛ اول اینکه صحت و درستی حقایق استخراج شده مورد ارزیابی قرار گیرد و جنبه دوم مربوط به ارزشمندی این حقایق می‌باشد. با توجه به استخراج خودکار حقایق، معیار سنجش صحت و ارزشمندی یکی از چالش‌ها در ارزیابی اطلاعات در استخراج آزاد می‌باشد. ارزیابی سیستم‌های استخراج اطلاعات با استفاده از افراد خبره و یا میزان کارایی کاربرد آنها در سیستم‌های دیگر مورد بررسی قرار می‌گیرد.

۲-۲-۲- شبکه واژگان

برای پردازش زبان طبیعی از منابعی مانند شبکه واژگان استفاده می‌شود. واژگان معنایی، نقش اساسی در سیستم‌های پردازش زبان طبیعی ایفا می‌کنند. شبکه واژگان یک پایگاه داده لغوی است که ابتدا برای زبان انگلیسی و با مدیریت George A. Miller توسعه پیدا کرد. اهداف اصلی ایجاد شبکه واژگان عبارتند از [23]:

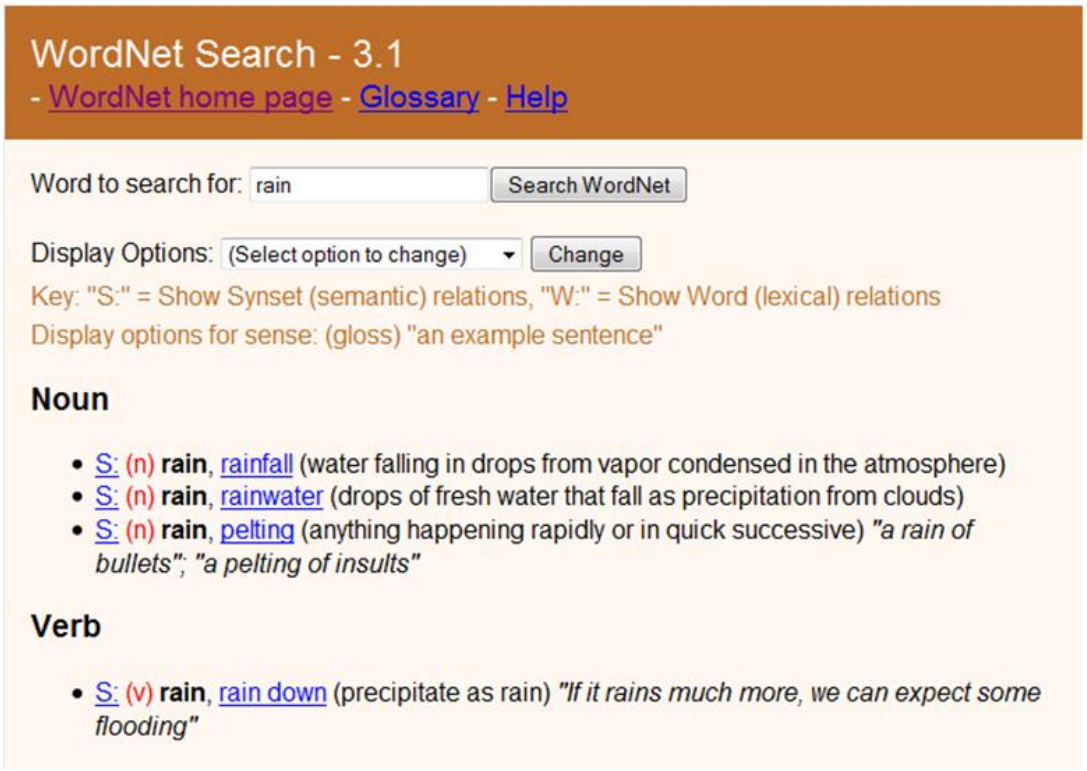
- ساخت ترکیبی از فرهنگ لغت و دایره المعارف که کاربردی‌تر باشد.
- پشتیبانی از آنالیز خودکار متون و کاربردهای هوش مصنوعی

البته بعدها کاربردهای دیگری نیز به این دو هدف اضافه شدند. شبکه واژگان، لغات را در مجموعه‌های هم‌خانواده قرار می‌دهد. هر کدام از این مجموعه‌های هم‌خانواده شامل لغاتی است که در

یک متن می‌توانند به جای یکدیگر استفاده شوند. در واقع هر مجموعه هم‌خانواده یک مفهوم خاص را بیان می‌کند. مجموعه‌های هم‌معنی به وسیله مفاهیم معنایی و روابط لغوی به یکدیگر مرتبط می-شوند. در نتیجه شبکه واژگان، شبکه‌ای است متشکل از لغات و مفاهیم که از نظر معنایی با یکدیگر ارتباط دارند [31].

ایجاد دستی شبکه واژگان، وقت‌گیر و هزینه‌بر می‌باشد و به دانش زبان‌شناسی نیاز دارد. برای ایجاد شبکه واژگان زبان‌های دیگر، روش‌های خودکار پیشنهاد شده است که از شبکه واژگان انگلیسی و منابع لغوی دیگر استفاده می‌کنند.

در شبکه واژگان ابتدا لغات در یکی از دسته‌های اسم، فعل، صفت، و قید تقسیم می‌شوند و لغات هرکدام از این دسته‌ها در مجموعه‌های هم‌خانواده قرار می‌گیرند. روابط معنایی متفاوتی در شبکه واژگان با مقوله‌های متفاوت (اسم، فعل، صفت، و قید) وجود دارد. این چهار دسته از لغات در مجموعه‌های هم‌خانواده شامل چندین لغت و یا ترکیبی از لغات است که به‌صورت مشترک برای بیان یک مفهوم خاص آورده می‌شوند. ممکن است برای هرکدام از این مجموعه‌های هم‌خانواده، یک توصیف کوتاه و همچنین جملات نمونه نیز ثبت شود. بسته به اینکه واژه به چه مقوله دستوری تعلق داشته‌باشد، روابط معنایی متفاوتی دارد [32]. در شکل (۲-۱) مجموعه‌های هم‌خانواده‌ی کلمه «rain» در شبکه واژگان انگلیسی نشان داده شده است.



شکل ۲-۱ : مجموعه‌های هم‌خانواده‌ی کلمه «rain» در شبکه واژگان انگلیسی

شبکه واژگان بیش از آنکه به طور مستقل به کار گرفته شود، در سایر پروژه‌ها به عنوان یک ابزار استفاده می‌شود. یکی از دلایل استقبال از شبکه واژگان، ساختار پایگاه داده آن است که باعث شده در محاسبات زبانی و همچنین پردازش زبان طبیعی به عنوان یک ابزار سودمند استفاده شود. در اکثر پروژه‌ها می‌توان از شبکه واژگان به صورت مستقیم برای کاربردهای مبتنی بر دانش، مخصوصا در پروژه‌های استخراج اطلاعات و بازیابی اطلاعات استفاده کرد [33].

RDF –۳–۲–۲

هدف اصلی در وب معنایی، توصیف اشیا (منابع[۱۲]) و روابط میان آنها است که برای بیان این اطلاعاتاز مدل داده RDF استفاده می‌شود. همانطور که در سیستم‌های اطلاعاتی، مدل داده رایج،

[۱۲] resource

مدل رابطه‌ای می‌باشد، در وب معنایی هم مدل داده استاندارد، مدل RDF است که یک مدل داده گرافیکی ساده است [34].

عناصر اصلی در مدل دادهRDF، سه‌گانه‌هایی به صورت (شی، صفت، مقدار) می‌باشند. در واقع هر سه‌گانه، یک عبارت مسندی[۱۳] را به صورت (نهاد[۱۴]، مسند، گزاره[۱۵]) نشان می‌دهد. هر بخش یک سه‌گانه، یک گره از گراف RDF را تشکیل می‌دهد. در حالت کلی یک سه‌گانه RDF به صورت نمادین زیر نمایش داده می‌شود:

$$t=(s,p,o) \in (U \cup B)\times U\times(U \cup B \cup L) \qquad (۱-۲)$$

در این‌جا U، شناسه یکتای منبع[۱۶](URI) است. هر منبع در وب معنایی، دارای یک شناسه یکتا می‌باشد که توسط آن شناخته می‌شود. یک URI ممکن است یک URL یا آدرس وب نیز باشد.در گراف RDF، بعضی از گره‌ها ممکن است ارجاع URI نداشته باشند که به آنها گره‌های تهی (B) گفته می‌شود.مقادیر و یا به عبارت دیگر گزاره‌ها در یک سه گانه RDF، می‌توانند از نوع مقدار ثابت[۱۷](L) هم باشند [35].

برای توصیف خصیصه‌ها و کلاس‌های منابع RDF و بیان سلسله مراتب بین آن‌ها از شمای RDF (RDFS) که در واقع یک لغت‌نامه است، استفاده می‌شود. با استفاده از زبان هستان‌شناسی وب (OWL) نیز می‌توان لغات و معنای بیشتری را برای توصیف منابع بیان کرد. اسناد وب معنایی خالص، اسنادی هستند که با گرامر RDF و یا به زبان OWL نوشته می‌شوند. علاوه بر این، گرامرهای RDF

---

[۱۳]predicate

[۱۴]subject

[۱۵]object

[۱۶]Uniform Resource Indicator

[۱۷] Literal

دیگری مانند Turtle، N-Triples و N-Quads نیز وجود دارد که از آن‌ها می‌توان برای بیان داده‌ها به صورت معنایی استفاده کرد.

۲-۳- روش‌های استخراج اطلاعات

همانطور که در مقدمه اشاره شد، روش‌های جمع‌آوری حقایق، روش‌های مختلفی را دنبال می‌کنند:

۱. مبتنی بر قاعده با پرس‌وجوی اعلانی برای ورودی‌های نیمه ساخت یافته مانند جداول وب مناسب است.
۲. مبتنی بر الگو
۳. مبتنی بر استدلال
۴. مبتنی بر پردازش زبان طبیعی
۵. مبتنی بر هستان‌شناسی

در ادامه این پنج روش و سیستم‌هایی که از این روش‌ها استفاده می‌کنند، مورد بحث و بررسی قرار می‌گیرند.

۲-۳-۱- روش مبتنی بر قاعده

در خیلی از موارد، صفحات وب معمولا از سیستم مدیریت محتوای دارای پایگاه داده تولید شده‌اند. در نتیجه امکان ایجاد و استنتاج خودکار برای استخراج حقیقت از سرفصل‌های HTML، جداول، لیست‌ها، فیلدهای فرم و عناصر دیگر نیمه ساخت یافته فراهم می‌شود. بعضی تکنیک‌های

یادگیری ماشین مانند HMM[۱۸] و کلاس‌بندی‌ها هم می‌توانند استفاده شوند، اما بنیاد کلی بر مجموعه قواعد استخراج مناسب، می‌باشد.

برای نمونه می‌توان از سیستم‌های اولیه Rapier و W4F toolkit و سیستم‌های اخیر Lixto، RoadRunner و SEAL نام برد. بیشتر کارهای اخیر در جمع‌آوری حقیقت مبتنی بر قاعده بر روی استخراج اطلاعات اعلانی[۱۹]DB-styleمی‌باشد [1].

استخراج حقایق مبتنی بر قاعده به صورت گسترده در استخراج اطلاعات منابعی مانند ویکی‌پدیا بخصوص برای استخراج جعبه اطلاعات[۲۰] استفاده می‌شود. پروژه‌های DBpedia، YAGO و Kylin/KOG از این روش بهره‌مند هستند. پروژه YAGO[9] بر اساس موجودیت‌ها و روابط ایجاد شده است. در YAGO، حقایق بصورت خودکار از ویکی‌پدیا و شبکه واژگان با استفاده از ترکیب روش‌های مبتنی بر قاعده و ابتکاری، استخراج می‌شوند. ارزیابی عملی نشان می‌دهد که دقت درستی حقایق حدود ۹۵٪ می‌باشد. YAGO بر اساس یک مدل منطقی است که تصمیم پذیر، قابل گسترش و سازگار با RDFS می‌باشد. YAGO یک هستان‌شناسی جدید با پوشش بالا[۲۱] همراه با کیفیت بالا می‌باشد [9].

---

[۱۸] Hidden Markov Model

[۱۹] declarative IE

[۲۰] infobox

[۲۱] high coverage

۲-۳-۲- روش مبتنی بر الگو

الگوهای متنی، برای استخراج حقیقت از اسناد غیرساخت‌یافته استفاده می‌شوند، اما برای ورودی‌های نیمه ساخت‌یافته مانند جداول وب یا لیست‌ها هم مناسب هستند. برای ایجاد الگوها، می‌توان از روش‌های دستی، یادگیری با ناظر[۲۲] و یادگیری خودناظر[۲۳] استفاده نمود.

در روش دستی، عبارات منظم یا قواعد، توسط خود انسان تهیه می‌شود. در نتیجه سیستم‌ها هزینه‌بر هستند و مقیاس‌پذیر و قابل حمل نیستند.

اگر حقایق کافی و مناسب برای یک رابطه موجود باشد، سیستم می‌تواند به صورت خودکار الگوهای متنی را تولید کند و بهترین آنها را انتخاب کند و همچنین اگر الگوهای خوبی موجود باشد، می‌توان حقایق مناسب بیشتری را به صورت خودکار تولید نمود [7]. برای جلوگیری از تولید حقایق نامناسب، الگوها مورد ارزیابی قرار می‌گیرند. وزن هر الگو با توجه به مدل‌های آماری محاسبه می‌شود. الگوهایی که حقایق نامناسب تولید کنند، وزن آنها به سمت صفر میل می‌کند و از روند استخراج اطلاعات حذف می‌شوند.

در متدهای باناظر، مجموعه آموزشی نمونه‌های برچسب خورده در دامنه خاص، به سیستم داده می‌شود. با استفاده از روش‌های یادگیری ماشین، الگوهای استخراج به صورت خودکار برای استخراج حقایق از متن تهیه می‌شوند. سیستم‌های DIPRE [36] ، Snowball [3, 36] و Meta-[1] Bootstrapping از این روش استفاده می‌کنند. با پیدایش چنین سیستم‌هایی، زحمت انسانی برای استخراج اطلاعات کمتر شده است. البته استخراج وابسته به روابط مشخصی می‌باشد و نیاز است که کاربر مجموعه نمونه رابطه مورد نظر یا یک مجموعه الگوهای استخراج دستی اولیه را برای شروع روند یادگیری آماده کند [2].

---

[22] Supervised Methods

[۲۳] Self-Supervised Methods

سیستم DIPRE با جستجوی داده‌های ورودی سعی می‌کند، الگوی مناسبی برای آن پیدا کند. هر داده با ویژگی‌های ترتیب وقوع آرگومان‌ها، آدرس وقوع، متن چپ، متن وسط و متن راست توصیف می‌شود. با این توصیف می‌توان الگوهای مختلف را در آدرس‌های مختلف وب و با انطباق متن‌های اطراف آرگومان استخراج کرد. مثلا الگوی

"ARG1''s headquarters in ARG2"

ممکن است در این فرایند استخراج شود. انتخاب نوشته‌های اطراف آرگومان‌ها با اشتراک‌گیری میان وقوع مختلف نمونه‌ها مشخص می‌شود. برای جلوگیری از خطاهای واضح، از یک عبارت با قاعده برای توصیف آرگومان‌ها، استفاده می‌شود [36].

سیستم SnowBall هم مانند DIPRE عمل می‌کند، البته قواعد استخراج با استفاده از شناسایی موجودیت‌های نام‌دار، قوی‌تر شده است. مثلا الگوی

"ORGANIZATION's headquarter in LOCATION"

به جای الگوی DIPRE استخراج می‌شود، و احتمال وقوع به هر یک از واژه‌های مربوط به الگو با استفاده از محاسبه تکرارهای مختلف نمونه‌ها اضافه می‌شود. حلقه تکرار این سیستم نیز مطابق دیگر نمونه‌های خود راه‌انداز، با جمع‌آوری داده مطمئن برای بهتر یاد گرفتن الگو، عمل می‌کند [3, 36].

سیستم‌های خودناظر، مانند now-ItAll [12] فقط با یک مجموعه کوچک از الگوهای استخراج مستقل از دامنه، برچسب گذاری مثال‌های آموزشی خودش، را انجام می‌دهد. سیستم Know-ItAll اولین سیستم استخراج صفحات وب بدون ناظر مستقل از دامنه و در اندازه بزرگ است. در سیستم‌های خود ناظر به جای استفاده از داده آموزشی برچسب خورده، خودش مثال‌ها را انتخاب می‌کند و برچسب‌گذاری می‌کند. سیستم‌های خودناظر جزء سیستم‌های بدون ناظر هستند، ولی برخلاف آنها از مثال‌های برچسب خورده استفاده می‌کند [12].

برای مثال، KnowItAll، از الگوهای استخراج عمومی مانند "<X> is a <Y>" برای شناسایی لیست اعضای کاندید X از کلاس Y استفاده می‌کند. وقتی این الگو برای کلاس Country بکار می‌رود، آن باید با یک جمله‌ای که شامل عبارت "X is a Country" باشد، تطبیق داده شود. بعد KnowItAll، بصورت متناوب، محاسبات آماری را توسط پرس‌وجوی موتورهای جستجو برای تشخیص نمونه‌هایی که شباهت بیشتری به اعضای کلاس دارند، بکار می‌برد. برای مثال، احتمال اینکه "Iran" نام یک کشور باشد، را تخمین می‌زند. KnowItAll عبارات تولید شده مرتبط با کلاس Country، را بکار می‌برد تا همبستگی، بین اسناد شامل کلمه "Iran" و آن‌هایی که شامل عبارت "countries such as" باشد، را محاسبه کند. بنابراین سیستم، می‌تواند "Iran"، "France" و "India"را به عنوان یک عضوی از کلاس Country برچسب‌گذاری کند. سیستمKnowItAll یک مجموعه الگوهای استخراج رابطه خاص، مانند "capital of <country>" برای استخراج کشورها یاد می‌گیرد [2, 12].

سیستم‌های خودناظر، KnowItAll، محدود به روابط خاص می‌باشند و برای هر رابطه، روند پر زحمت راه‌اندازی لازم است و مجموعه روابط باید توسط انسان مشخص شود. ولی سیستم اغلب، با بعضی مفاهیم و روابط پیش‌بینی نشده، در زمان پردازش متن مواجه می‌شود.

۲-۳-۳- روش مبتنی بر استدلال

در روش مبتنی بر استدلال، یک عمل اضافی استدلال بر روی مجموعه فرضیات حقایق جمع‌آوری شده، انجام می‌شود. در نتیجه حقایق جدیدی به پایگاه دانش اضافه می‌شود. در این قسمت سازگاری حقایق در پایگاه داده بررسی می‌شود، یعنی بتوان حقایقی نادرست را با توجه به نقص محدودیت‌ها تشخیص داد، و آن‌ها را حذف نمود. برای مثال، نوع موجودیت‌های شرکت کننده در یک رابطه بررسی می‌شوند، برای مثال در رابطه تدریس باید موجودیت اول انسان و به صورت دقیق‌تر یک مدرس، و موجودیت دوم یک مرکز آموزشی باشد، تا رابطه تدریس تایید شود.

۲-۳-۴- روش مبتنی بر پردازش زبان طبیعی

تکنیک‌های پردازش زبان طبیعی مانند برچسب‌گذاری POS، تجزیه وابستگی و برچسب‌گذاری نقش معنایی[۲۴] برای استخراج اطلاعات در اسناد استفاده می‌شوند. این تکنیک‌هابرای ایجاد رابطه میان عبارات و عناصر جمله، استفاده می‌شوند. همچنین می‌توان قوانین استخراج بر اساس محدودیت‌های معنایی و نحوی بدست آورد. از این قوانین می‌توان در استخراج اطلاعات از یک سند استفاده نمود [15, 16].

ابزارهایی مانند RAPIER[16] و WHISK[15] از روش مبتنی بر NLP بهره‌می‌برند. در استخراج اطلاعات می‌توان هم از NLP سطحی و هم از روش NLP عمقی[۲۵] استفاده کرد. آنالیز نحوی سطحی برای استخراج روابط یک دامنه خاص و آنالیز عمقی برای استخراج روابط بر اساس فعل بکار می‌رود [15, 17]. برای تشخیص روابط در ساختارهای پیچیده از DNLP استفاده می‌شود. بعضی سیستم‌ها مانند WHIES[17] هم از SNLP و هم از DNLP استفاده می‌کنند. ارزیابی سیستم‌ها نشان می‌دهد که DNLP باعث افزایش کارایی استخراج اطلاعات می‌شود [17].

۲-۳-۵- روش مبتنی بر هستان‌شناسی

استخراج اطلاعات مبتنی بر هستان‌شناسی[۲۶]OBIE، یک روند شناسایی مفاهیم، خصیصه‌ها و روابط بیان شده در هستان‌شناسی، از متن و منابع دیگر است. آنتولوژی شامل مفاهیم تنظیم شده در سلسله مراتب کلاس و زیرکلاس، روابط بین مفاهیم و خصیصه‌ها می‌باشد [18, 21, 22]. اختلاف روش‌های استخراج اطلاعات دیگر و OBIE در نحوه استفاده از هستان‌شناسیاست. هستان‌شناسی حتی می‌تواند در استنتاج سهیم باشد. همچنین موجودیت‌ها می‌توانند با بررسی اسناد، توضیحات خود

---

[۲۴] semantic role labeller(SRL)

[۲۵] deep NLP (DNLP)

[۲۶] Ontology-based Information Extraction

را در طی روند استخراج اطلاعات غنی‌تر کنند.اگر نمونه‌های مناسبی در هستان‌شناسی موجود باشد، شناسایی آنها در متن ساده می‌شود، البته در این جا استفاده از هستان‌شناسی بیش از یک فرهنگ لغت می‌باشد [19].

مزیت سیستم‌های OBIE نسبت به سیستم‌های سنتی استخراج اطلاعات این است که خروجی به یک هستان‌شناسی لینک داده می‌شود. این ارتباط موجب استخراج اطلاعات بیشتری درباره متن، با استفاده ازاطلاعات رابطه‌ای یا استدلال می‌شود [21].

سیستم‌های مختلفی از قبیل KIM [19]، SemTag[21] و Artequakt[18] را می‌توان به عنوان نمونه‌های سیستم‌های مبتنی بر هستان‌شناسی نام برد. سیستم Artequakt یک سیستم استخراج اطلاعات است که روابط موجودیت را با استفاده از تعاریف رابطه هستان‌شناسی و اطلاعات لغوی شناسایی می‌کند.

در Artequakt یک هستان‌شناسی با شبکه واژگان، GATE[۲۷] و یک تشخیص دهنده موجودیت، برای شناسایی موجودیت‌ها و تشخیص روابط مابین آن‌ها استفاده می‌شود. GATE یک معماری برای مهندسی زبان است که توسط دانشگاه Sheffield توسعه یافته است، و شامل ابزاری مناسب برای پردازش زبان طبیعی می‌باشد. در مرحله اول، اسناد مربوطه، جستجو می‌شوند و بعد سیستم استخراج اطلاعات، سند منتخب را به پاراگراف و سپس جملات تقسیم می‌کند و هر قسمت را از نظر معنایی و نحوی مورد بررسی قرار می‌دهد. با استفاده از تجزیه‌گر تحلیل نحوی و با استفاده از تحلیل معنایی اجزای مهم جمله (فاعل، فعل، مفعول) مشخص می‌کند و موجودیت‌های نام‌دار را با استفاده از GATE و شبکه واژگان شناسایی می‌کند. سیستم همچنین با استفاده از GATE مسئله ارجاعات را حل می‌کند [18].

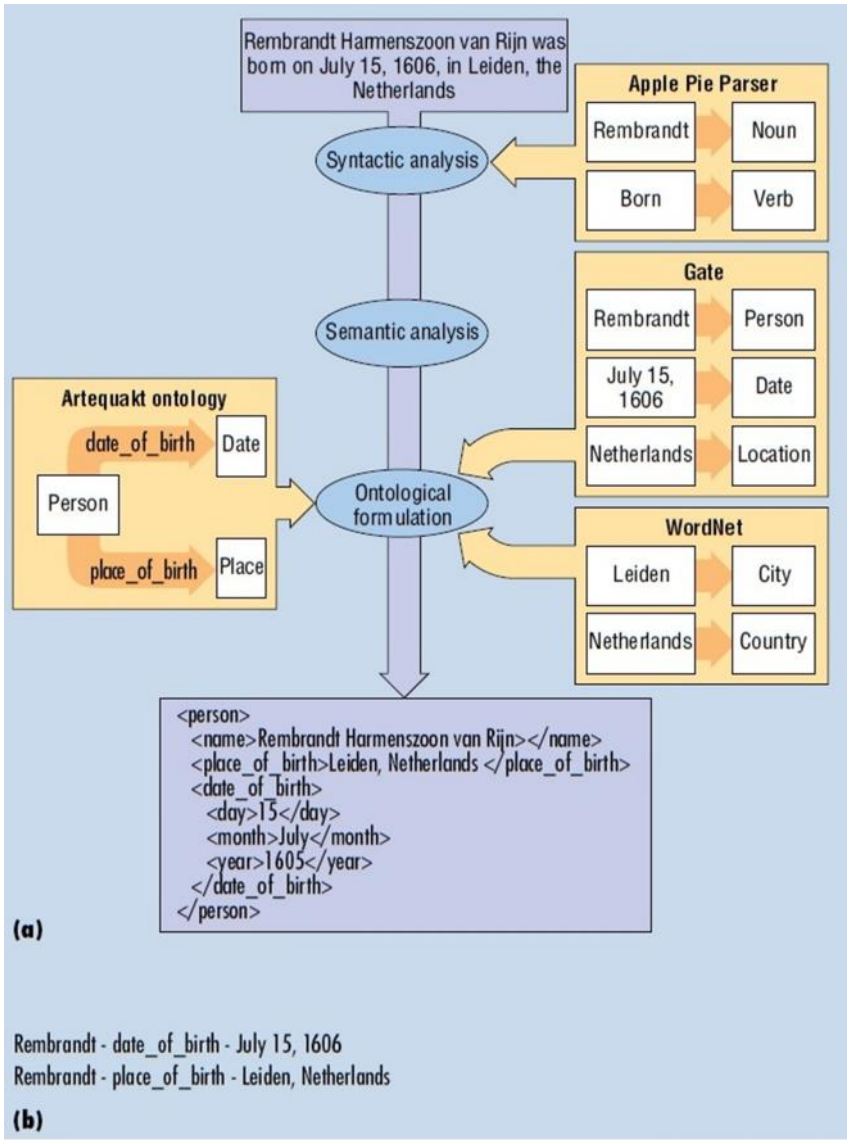
---

[۲۷] General Architecturefor Text Engineering

برای استخراج روابط دودویی بین یک جفت موجودیت، به دانش معنایی برای یک دامنه خاص نیاز است که می‌توان از هستان‌شناسی برای تشخیص روابط لازم و مورد انتظار بین موجودیت‌ها استفاده نمود. همچنین از شبکه واژگان (برای مترادف‌ها، hypernymها و hyponymها) برای کاهش اختلاف لغوی میان روابط شناخته شده در هستان‌شناسی و متن استفاده می‌شود.

برای مثال در شکل (۲-۲)، رابطه date_of_birth به مفهوم birth بر طبق شبکه واژگان که چهار معنی اسمی و یک معنی فعلی دارد، نگاشت می‌شود. سیستم، اولین معنی اسمی را انتخاب می‌کند، زیرا یکی از hypernyms، دوره زمانی، date را بعنوان hyponym دارد. مترادف‌های کسب شده برای معنی فعلی شامل give birth و bear است.

شبکه واژگان و GATE مشخص می‌کنند که Rembrandt Harmenszoon van Rijn نام شخص، 15 July 1606 تاریخ، Leiden و Netherlands مکان هستند. و نتیجه استخراج، دو رابطه date_of_birth و place_of_birth می‌باشد.



شکل ۲-۲:یک نمونه استخراج دانش مبتنی بر هستان‌شناسی [18]

۲-۳-۶- مقایسه و ارزیابی

در جدول (۲-۱) مقایسه‌ای از ویژگی‌های مربوط به تعدادی ازسیستم‌های استخراج اطلاعات ارائه شده است. این سیستم‌ها با توجه به استفاده از روش‌های استخراج اطلاعات بررسی شده‌اند.

جدول ۲-۱: مقایسه سیستم‌های استخراج اطلاعات

سیستمو پروژه‌های استخراج اطلاعات	YAGO [9]	Snowball [36]	KnowItAll [12]	WHIES [17]	h-TechSigh [20]	[18] Artequakt
نیازمندی‌ها	شبکه واژگان	مجموعه داده ورودی برای آموزش	تعدادی الگو برای شروع کار، استفاده از موتورهای جستجو	گرامرهای مبتنی بر الگو	هستان‌شناسی از دامنه مورد نظر	هستان‌شناسی از دامنه مورد نظر، اطلاعات لغوی شبکه واژگان
روش استخراج رابطه	مبتنی بر قاعده	مبتنی بر الگو با روش یادگیری با ناظر	مبتنی بر الگو روش یادگیری خودناظر مستقل از دامنه	مبتنی بر NLP	مبتنی بر هستان‌شناسی	مبتنی بر هستان‌شناسی
چالش‌ها	محدود به روابط خاص، نیاز به قواعد استخراج	محدود به روابط خاص، عدم مقیاس پذیر در سطح وب	وابسته به روابط، کند بودن	کند بودن به علت نیاز به DNLP	محدود به دامنه خاص	محدود به رابطه خاص
نتایج ارزیابی	با دقت بالای ۹۵٪	دقت ۸۵٪ با فراخوانی ۶۸٪	مقیاس پذیر، در ۴۰۰ میلیون صفحه با دقت ۹۰٪	با ترکیب DNLP و SNLP فراخوانی ۹۴٪	-	مقیاس‌پذیر
تکنیک‌های مورد استفاده	غنی سازی با استنتاج	ارزیابی‌های آماری برای بهبود الگوها	ارزیابی‌های آماری برای بهبود	DNLP (SRL)	GATE	GATE

۲-۴- استخراج آزاد اطلاعات

در روش استخراج آزاد اطلاعات همه روابط بامعنا استخراج می‌شود. در این روش، تعداد روابط نامحدود می‌باشد و بیشتر از روابط متعارف از پیش تعیین شده، می‌باشد؛ روابط به یک دامنه خاص محدود نمی‌شود؛ و همه عبارات فعلی می‌توانند کاندید روابط بین موجودیت‌ها باشند [37, 38]. در جدول (۲-۲) تفاوت میان استخراج سنتی اطلاعات با استخراج آزاد اطلاعات، نشان داده شده است:

جدول ۲-۲: تفاوت بین استخراج سنتی و آزاد اطلاعات [2]

	IE آزاد	IE سنتی
ورودی	متن + روش‌های مستقل از دامنه	متن+ داده برچسب خورده[۲۸]
روابط	کشف خودکار	مشخص
پیچیدگی	$O(D)$ D اسناد	$O(D * R)$ D اسناد، R روابط

در ادامه برخی سیستم‌های استخراج آزاد و چالش‌های آنها مورد بررسی قرار می‌گیرد.

۲-۴-۱- سیستم TEXTRUNNER

سیستم TextRunner[39]، جزء اولین سیستم‌های استخراج آزاد اطلاعات می‌باشد، که می‌تواند مجموعه بزرگی از تاپل‌های رابطه‌ای را بدون نیاز به ورودی انسان استخراج کند. سیستم TextRunner یک مثال از OIE با مقیاس‌پذیری بالا می‌باشد؛ و از ۱۲۰ میلیون صفحه وب، ۵۰۰ میلیون حقیقت سه‌گانه با میانگین دقت ۷۵٪ استخراج کرده است [2].

[۲۸] Labeled Data

OIE چالش‌های جدید و مهمی را برای سیستم‌های استخراج اطلاعات ارائه می‌کند [39, 40, 4]:

- پیچیدگی استخراج خودکارروابط که سیستم‌های استخراج اطلاعات سنتی ورودی‌های برچسب خورده دستی استفاده می‌کنند.
- ناهمگنی کالبد[۲۹] متن بر روی وب که دقت ابزاری مانند تجزیه‌گرها و برچسب‌گذاری‌های موجودیت نام‌دار[۳۰]را به علت تفاوت متن از داده آموزشی ابزار، کاهش می‌دهد.
- مقیاس‌پذیری و کارایی که سیستم‌های OIE به گذر سریع و یک‌باره در سطح داده که می‌تواند مجموعه‌هایی از اسناد عظیم باشد، محدود کرده است [38].

سیستم TextRunner، ابتدا با اعمال تعدادی قواعد روی داده‌ها، برای خود تعدادی نمونه‌ی صحیح فراهم می‌کند. برای استخراج اطلاعات از وب، استفاده از تجزیه نحوی جمله‌ها بسیار زمان‌بر است؛ در نتیجه ابتدا با چند قاعده تولید شده دستی، از روی درخت نحوی جمله‌ها، نمونه‌های آموزشی مناسبی را برای خود فراهم می‌کند. البته با اجرای یک‌باره می‌تواند به داده‌های آموزشی مورد نظر خود دست یابد. در این سیستم رابطه‌های سه‌گانه، شامل یک رابطه و دو موجودیت استخراج می‌شود. پس هدف یافتن جفت عبارات اسمی است که خیلی از هم دور نیستند، و با توجه به محدودیت‌ها، و اعمال یک رده‌بندی کننده[۳۱] مشخص می‌شود که آیا یک رابطه استخراج می‌شود یا نه. برای هر دو عبارت اسمی در جمله، باید مشخص شود که کدام واژه‌های جمله، جزئی از رابطه میان آن دو هستند [39, 40].

---

[۲۹] Corpus Heterogeneity

[۳۰] named entity

[۳۱] classifier

حال با استفاده از محدودیت‌های ابتکاری، به صورت خودکار، مجموعه جملات تجزیه شده به عنوان استخراج‌های درست و نادرست (مثال‌های مثبت و منفی) برچسب می‌خورند. رده‌بندی کننده بر اساس این مثال‌ها، با بکار بردن ویژگی‌هایی مانند برچسب‌های POS بر روی کلمات رابطه، آموزش می‌بیند. سپس رده‌بندی کننده می‌تواند تصمیم بگیرد که آیا یک ترتیبی از کلمات برچسب خورده POS یک استخراج مناسب با دقت بالا است [4, 39].

سیستم TextRunner، هیچ رابطه از پیش تعریف شده‌ای ندارد، به همین علت، ممکن است خیلی از رشته‌های مختلف، یک رابطه را ارائه دهند. سیستم RESOLVER[39]، یک خوشه‌بندی بدون ناظر، برای ایجاد مجموعه‌های موجودیت‌ها و روابط مترادف انجام می‌دهد، به عبارت دیگر، برای مشخص کردن احتمال هم‌مرجعی هر جفت رشته استفاده می‌کند.

سیستم TextRunner، نمونه‌هایی را که با قاعده استخراج شده‌اند، را یاد می‌گیرد. در نتیجه به تجزیه‌ی نحوی جمله در زمان استخراج اطلاعات بی‌نیاز است؛ و خطای تجزیه، منجر به خطا در استخراج اطلاعات نخواهد شد و روند استخراج با سرعت بیشتری انجام می‌شود.

در یک مقایسه سیستم‌های TextRunner و KnowItAll بر روی ۹ میلیون صفحه وب، با در نظر گرفتن فقط ۱۰ رابطه، نسبت خطای TextRunner نسبت به KnowItAll کمتر است، و آرگومان‌های مناسب‌تری برای روابط شناسایی کرده است [39].

۲-۴-۲- سیستم WOE

سیستم WOE[7] با توجه به ایده TextRunner از ویکی‌پدیا برای داده‌های آموزشی رده‌بندی کننده، استفاده می‌کند. برای ایجاد داده‌های آموزشی از داده‌های ساخت‌یافته بهره‌مند می‌شود و این داده‌ها را همراه با عنوان مقاله، در جملات جستجو می‌کند. البته با توجه به نام‌های مختلف برای یک مفهوم در ویکی‌پدیا، این سیستم از ارجاع‌های مختلف به یک صفحه استفاده می‌کند تا فقط به عنوان

مقاله محدود نشود. WOE^{POS} معادل سیستم TextRunner، که تنها از ویژگی‌های برچسب‌گذاری POS و WOS^{parse} از داده‌های مربوط به تجزیه وابستگی جمله بهره می‌گیرد [4, 7, 8].

۲-۴-۳- سیستم StatSnowball

سیستم StatSnowball[3]، می‌تواند هم استخراج سنتی و هم استخراج آزاد اطلاعات را انجام دهد. آزمایش‌های تجربی نشان می‌دهد که StatSnowball، به طور چشم‌گیری فراخوانی بالایی را بدون قربانی کردن دقت بدست می‌آورد. در ابتدا استخراج روابط معنایی از پیش تعریف شده، بین جفت موجودیت‌ها در متن با روش‌های یادگیری باناظر انجام می‌گرفت، اما برچسب‌گذاری مثال‌های آموزشی سخت است و در سطح وب مقیاس‌پذیر نیست [3].

سیستم Snowball[36]، یک مجموعه کوچک از تاپل‌ها، را به عنوان ورودی می‌گیرد و دوگانه الگو-موجودیت[۳۲]را بکار می‌برد و مکررا الگوهای استخراج را تولید می‌کند و تاپل‌های رابطه جدید را شناسایی می‌کند. ارزیابی الگوها و تاپل‌ها یکی از اجزای کلیدی برای اطمینان از جلوگیری از انحراف سیستم به سمت اشتباه است. اگرچه معماری خود راه‌اندازها امیدبخش است ولی برای اندازه وب، به علت دو محدودیت، مناسب نیست. در این سیستم استخراج نوع مشخصی از روابط انجام می‌شود و الگوهای استخراج بر اساس تطبیق کلمه کلیدی سخت[۳۳] هستند و اگرچه دقت بالا می‌رود ولی فراخوانی کاهش می‌یابد. همچنین دارای ارزیابی خوبی نیست، برای نمونه دقت و معیارهای انتخاب الگو به صورت مستقیم با الگوهای عمومی قابل توافق نیستند، ممکن است خیلی از تاپل‌های استخراج شده توسط الگوی عمومی برای رابطه مورد نظر نباشد و برای رابطه دیگری باشد. در این حالت اندازه اطمینان کوچک می‌شود و این معیار برای انتخاب الگوها مناسب نیست[3, 36] .

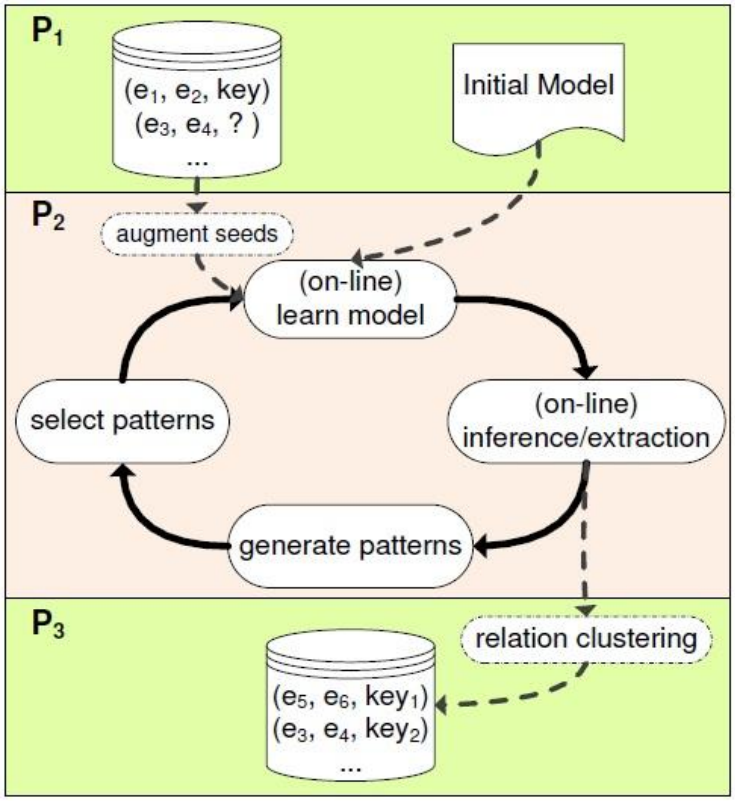
---

[۳۲] pattern-entity duality

[۳۳] strict keyword-matching

سیستم StatSnowball، می‌تواند هم استخراج‌های سنتی مانند Snowball، برای روابط از پیش تعریف شده و هم استخراج آزاد اطلاعات برای شناسایی انواع روابط انجام دهد. اگر چه سیستم‌های OIE موجود خودناظر هستند؛ و به یک مجموعه ویژگی‌های منتخب انسانی برای یادگیری بهتر استخراج‌گر نیاز دارند، در مقابل StatSnowball، به صورت خودکار، الگوهای استخراج را تولید و استخراج می‌کند. هر چند سیستم‌های OIE، به تکنیک‌های تجزیه زبانی عمقی و گران برای برچسب‌گذاری نمونه‌های آموزشی احتیاج دارد ولی StatSnowball از تکنیک‌های تجزیه سطحی و ارزان‌تر برای تولید الگو استفاده می‌کند. سیستمStatSnowball به راحتی قابل گسترش است. در پیاده‌سازی کنونی آن استراتژی یکسانی مانند Snowball، برای شناسایی جداگانه موجودیت‌های نام‌دار و روابط موجودیت در نظر گرفته شده است، ولی به راحتی می‌تواند این دو بخش را به یک مدل احتمالی ترکیب کند و کارایی بالاتری را کسب کند[3] .

در شکل (۲-۳) معماری StatSnowball، که به سه بخش تقسیم شده است، نمایش داده شده است. اولین قسمت، P1، ورودی می‌باشد که شامل یک مجموعه از نمونه‌ها و یک مدل اولیه است. نمونه‌ها نیازی به کلمات کلیدی برای مشخص کردن روابط ندارند. بنابراین دو نوع نمونه موجود است، نمونه‌هایی با کلمات کلیدی رابطه مانند (e1,e2,key) یا نمونه‌هایی بدون کلمات کلیدی رابطه مانند (e3,e4,?). اگر مدل اولیه خالی باشد، از نمونه‌ها برای تولید الگوهای استخراج در شروع فرایند استفاده می‌شود.



شکل ۲-۳ معماری سیستم StatSnowball [3]

دومین بخش معماری سیستم، P2، مدل استخراج آماری است، که در ابتدا ورودی‌های بخش P1 را می‌گیرد، و برای آموزش استخراج‌گر استفاده می‌کند. مدل یادگیری برای استخراج تاپل‌های رابطه جدید بر روی متن استفاده می‌شود. سپس الگوهای استخراج را با توجه به تاپل‌های رابطه شناخته شده اخیر، تولید می‌کند؛ و مدل دوباره آموزش می‌بینند. یک الگوی خوب و مناسب باید بتواند یک تعادل بین دو معیار اختصاصی[۳۴] و پوشش[۳۵] برقرار کند. اختصاصی به این معنی است که الگو بتواند تاپل‌های رابطه را با کیفیت بالا شناسایی کند، در حالیکه پوشش نمایش دهنده این است که الگو بتواند از لحاظ آماری تعداد زیادی تاپل خوب را پیدا کند. در Snowball، الگوهای تولید شده، بر

[۳۴] specificity

[۳۵] coverage

اساس تطبیق کلمه کلیدی هستند، این الگوهای قوی دارای دقت بسیار بالایی هستند ولی پوشش آنها بسیار پایین است.

در آخر هم الگوهای مناسب انتخاب می‌شوند، با توجه به مدل‌های احتمالی الگوهایی که وزن آن‌ها به سمت صفر میل کند، حذف می‌شوند. یک قسمت اختیاری هم تقویت نمونه‌ها است که می‌تواند نمونه‌های بیشتری را پیدا کند و کیفیت نمونه‌های آموزشی را بالا ببرد.

سومین بخش معماری P3، برای انجام استخراج آزاد اطلاعات پیکربندی شده است. نتایج استخراج در بخش P2، تاپل‌های رابطه عمومی هستند. برای خوانایی بیشتر نتایج، روش‌های خوشه‌بندی برای گروه‌بندی تاپل‌های رابطه و انتساب کلمه کلیدی‌ها به آنها اعمال می‌شود. کلمه کلیدی نمونه‌ها در این قسمت پر می‌شود. سیستم StatSnowball، جفت موجودیت‌های مرتبط را شناسایی می‌کند و کلمات کلیدی را برای روابط مشخص می‌کند [3] .

۲-۴-۴- سیستم O-CRF

سیستم O-CRF[41]، با استفاده از CRF[۳۶]، می‌تواند روابط مختلفی را با دقت ۸۸,۳٪ و فراخوانی ۴۵,۲٪ استخراج نماید. بیشتر روابط با استفاده از یک مجموعه الگوی نحوی- لغوی مستقل از رابطه توصیف می‌شوند. در جدول (۲-۳) یک طبقه‌بندی از روابط دودویی نشان داده شده است که نزدیک به ۹۵٪ جملات متعلق به یکی از این هشت تقسیم‌بندی می‌باشد.

[۳۶]Conditional RandomFields

جدول ۲-۳: طبقه‌بندی روابط دودویی [41]

Relative Frequency	Category	Simplified Lexico-Syntactic Pattern
37.8	Verb	E_1 Verb E_2 *X established Y*
22.8	Noun+Prep	E_1 NP Prep E_2 *X settlement with Y*
16.0	Verb+Prep	E_1 Verb Prep E_2 *X moved to Y*
9.4	Infinitive	E_1 to Verb E_2 *X plans to acquire Y*
5.2	Modifier	E_1 Verb E_2 Noun *X is Y winner*
1.8	Coordinate_n	E_1 (and|,|-|:) E_2 NP *X-Y deal*
1.0	Coordinate_v	E_1 (and|,) E_2 Verb *X , Y merge*
0.8	Appositive	E_1 NP (:|,)? E_2 *X hometown : Y*

هر چند الگوهای نمایش داده شده در جدول، با حذف شرایط دقیق، بسیار ساده شده‌اند، ولی استخراج‌های درستی را نتیجه می‌دهد. برای نمونه در حالیکه خیلی از روابط با فعل، مشخص می‌شوند، اما با توجه به مشخصات متن لازم است تا مشخص شود که فعل موجود در بین دو موجودیت واقعا رابطه بین آن‌ها را مشخص می‌کند. سیستم O-CRF با استفاده از میدان تصادفی شرطی CRF، می‌تواند چگونگی تعریف روابط دودویی در زبان انگلیسی را آموزش ببینند. با استفاده از زنجیره خطی CRF می‌توان کارهای پردازش متن شامل تشخیص موجودیت نام‌دار، برچسب‌گذاری POS، تقسیم‌بندی کلمات، شناسایی نقش معنایی و استخراج روابط انجام داد [41].

سیستم O-CRF، فقط روابطی که به صورت واضح در متن ظاهر شده‌اند را استخراج می‌کند و روابط ضمنی باید از روابط استخراج شده، استنتاج شود. روابط باید بین نام‌های موجودیت در یک جمله قرار داشته باشد.

۲-۴-۵- سیستم ReVerb

در سیستم‌های OIE قبلی، دو مشکل استخراج‌های بی‌معنی[۳۷] و استخراج‌های خالی از اطلاعات مفید[۳۸] وجود دارد. استخراج‌های بی‌معنی، حالت‌هایی هستند که عبارات رابطه‌ای استخراج شده، تفسیر بامعنایی ندارد. استخراج‌های بی‌معنی، وقتی اتفاق می‌افتد که استخراج‌گر در مورد هر کلمه تصمیم بگیرد که در عبارت رابطه باشد یا نه، و اغلب عبارات رابطه‌ای غیرقابل فهم را نتیجه می‌دهد. دومین مشکل استخراج‌های خالی از اطلاعات مفید می‌باشد که اطلاعات حیاتی حذف می‌شود [4, 38].

سیستم ReVerb، برای رفع مشکلات ذکر شده، از محدودیت‌های نحوی و محدودیت‌های لغوی استفاده کرده است. بر اساس محدودیت‌های نحوی، عبارات رابطه‌ای باید با الگوی برچسب POS شکل (۲-۴) مطابق باشد. این الگو، عبارات رابطه‌ای را محدود به عبارت فعلی ساده یا عبارت فعلی به همراه حرف اضافه یا فعل به همراه یک یا چند اسم، صفت و مختوم به حرف اضافه می‌کند. اگر چندین تطبیق در جمله امکان داشته باشد، بلندترین تطابق انتخاب می‌شود [4].

```
V | VP | VWP
V = verb particle? adv?
W = (noun | adj | adv | pron | det)
P = (prep | particle | inf. marker)
```

شکل ۲-۴ : عبارات منظم مبتنی بر POS برای استخراج رابطه [4]

محدودیت‌های نحوی اگرچه موجب بالا رفتن دقت می‌شود ولی فراخوانی را کاهش می‌دهد. سیستم ReVerb با استفاده از یک دیکشنری بزرگ از عبارات رابطه‌ای، محدودیت‌های لغوی را اعمال می‌کندو از عبارات رابطه‌ای، فعل‌های کمکی، قیدها و ... را حذف می‌کند. فرایند استخراج با یافتن

---

[۳۷] incoherent extractions

[۳۸] uninformative extractions

فعل‌ها آغاز می‌شود و رابطه متناسب با هر فعل استخراج می‌شود. در مرحله بعد آرگومان‌های رابطه، با استفاده از عبارات اسمی سمت چپ و راست رابطه، مشخص می‌شوند. البته سیستم برای زبان انگلیسی طراحی شده است [4, 7, 8].

یکی از مشکلات سیستمReVerb، شناسایی نادرست آرگومان‌ها می‌باشد، یعنی تقریبا ۶۵٪ خطاهای ReVerb مربوط به عبارات رابطه‌ای صحیح با آرگومان‌های نامناسب است. که یک بخش یادگیری آرگومان به نام ARGLEARNER برای کاهش خطاها اضافه می‌شود. با توجه به اینکه مجموعه آموزشی بزرگی برای سیستم‌های OIE وجود ندارد، برای تولید داده آموزشی برای سیستم یادگیری، از برچسب‌گذاری نقش معنایی SRL استفاده می‌شود. جزئیات بیشتر در مقاله [7] توضیح داده شده است. سیستم ناشی از ترکیب استخراج رابطه ReVerb با استخراج آرگومان ARGLEARNER، R2A2 نامیده شده است [7]. سیستم OLLIE [8] از مجموعه داده ReVerb برای برای ایجاد الگوها خود استفاده می‌کند. سپس این الگوها را بر تجزیه وابستگی جملات اعمال می‌کند.

۲-۴-۶- استخراج آزاد اطلاعات مبتنی بر عبارت

استخراج اطلاعات مبتنی بر عبارت ClausIE[6][۳۹]، از دانش زبان‌شناسی درباره گرامر زبان انگلیسی برای تشخیص عبارات در جمله ورودی بهره می‌برد و سپس نوع هر عبارت را بر طبق تابع گرامری مشخص می‌کند. بر این اساس، ClausIE می‌تواند استخراج با دقت بالا را انجام دهد. سیستم ClausIE مبتنی بر تجزیه وابستگی و یک مجموعه کوچک لغوی مستقل از دامنه مانند افعال copular، جمله به جمله بدون هیچ پس‌پردازشی (برای حذف استخراج‌ها با دقت پایین) عمل می‌کند و به هیچ داده آموزشی نیاز ندارد و دارای دقت و فراخوانی بالایی نسبت به سیستم‌های دیگر می‌باشد.

[۳۹] Clause-Based Open Information Extraction

بعد از تشخیص عبارات، یک یا چند گزاره از هر عبارت با توجه به نوع آن، استخراج می‌شود. خطاهای استخراج در ClausIE به علت درخت‌های تجزیه نادرست است. البته می‌توان از تکنیک‌های پس‌پردازش مانند تکنیک‌های آماری در ReVerb برای افزایش دقت استفاده کرد. به طور چشم‌گیری آهسته‌تر از همه تکنیک‌های بالا می‌باشد؛ اما استخراج‌هایی با کیفیت بالا تولید می‌کند که حتی می‌تواند به عنوان داده آموزشی برای سیستم‌هایی مانند R2A2 مورد استفاده قرار گیرد.

در ClausIE نوع عبارات و افعال از نظر لازم و متعدی و یا افعال ترکیبی و کمکی و ... بررسی می‌شود. اطلاعات در مورد عبارت از طریق تجزیه وابستگی، و اطلاعات دو مورد انواع فعل از یک مجموعه کوچک لغوی مستقل از دامنه فراهم می‌شود. بدین وسیله با اطلاعات بدست آمده، می‌توان نوع عبارت را مشخص کرد. چالش‌های این سیستم وابستگی به گرامر و الگوهای مشخص برای عبارات زبان انگلیسی، استخراج موارد غیرحقیقی (واقعیت‌های اشتباه)، هزینه زمانی بالای آن می‌باشد.

در شکل (۲-۵) مثالی از استخراج حقایق در بعضی سیستم‌های استخراج آزاد اطلاعات نمایش داده شده است.

۲-۴-۷- ارزیابی و مقایسه

در جدول (۲-۴) مقایسه‌ای از ویژگی‌های مربوط به تعدادی از سیستم‌های استخراج آزاد اطلاعات ارائه می‌شود.

System	#	Proposition	Label
		Reverb dataset	
The principal opposition parties boycotted the polls after accusations of vote-rigging , and the only other name on the ballot was a little-known challenger from a marginal political party.			
Reverb	R_1:	(“*The principal opposition parties*”, “*boycotted*”, “*the polls*”)	Correct
	R_2:	(“*the ballot*”, “*was*”, “*a little-known challenger*”)	Incorrect
TextRunner	R_3:	(“*The principal opposition parties*”, “*boycotted*”, “*the polls*”)	Correct
WOE	R_4:	(“*The principal opposition parties*”, “*boycotted*”, “*the polls*”)	Correct
	R_5:	(“*the only other name*”, “*was*”, “*a little-known challenger*”)	Incorrect
OLLIE	R_6:	(“*The principal opposition parties*”, “*boycotted*”, “*the polls*”)	Correct
	R_7:	(“*The principal opposition parties*”, “*boycotted the polls after*”, “*accusations of vote-rigging*”)	Correct
	R_8:	(“*The principal opposition parties*”, “*was*”, “*a little-known challenger*”)	Incorrect
	R_9:	(“*The principal opposition parties*”, “*was a little-known challenger from*”, “*a marginal political party*”)	Incorrect
	R_{10}:	(“*the polls*”, “*be boycotted after*”, “*accusations of vote-rigging*”)	Correct
	R_{11}:	(“*the only other name*”, “*was*”, “*a little-known challenger*”)	Incorrect
	R_{12}:	(“*the only other name*”, “*was a little-known challenger from*”, “*a marginal political party*”)	Incorrect
	R_{13}:	(“*the only other name*”, “*boycotted*”, “*the polls*”)	Incorrect
	R_{14}:	(“*the only other name*”, “*boycotted the polls after*”, “*accusations of vote-rigging*”)	Incorrect
	R_{15}:	(“*a little-known challenger*”, “*be the only other name on*”, “*the ballot*”)	Correct
	R_{16}:	(“*only*”, “*be other name on*”, “*the ballot*”)	Incorrect
	R_{17}:	(“*other*”, “*be name on*”, “*the ballot*”)	Incorrect
ClausIE	R_{18}:	(“*The principal opposition parties*”, “*boycotted*”, “*the polls*”)	Correct (red.)
	R_{19}:	(“*The principal opposition parties*”, “*boycotted*”, “*the polls after accusations of vote-rigging*”)	Correct
	R_{20}:	(“*the only other name on the ballot*”, “*was*”, “*a little-known challenger*”)	Correct (red.)
	R_{21}:	(“*the only other name on the ballot*”, “*was*”, “*a little-known challenger from a marginal political party*”)	Correct

شکل ۲ -۵: مثالی از استخراج حقایق در برخی سیستم‌های استخراج آزاد اطلاعات [6]

جدول ۲-۴: مقایسه سیستم‌های استخراج آزاد اطلاعات

سیستم	TextRunner [39]	WOE [7]	StatSnowball [3]	O-CRF [41]	ReVerb [4]	R2A2 [7]	ClausIE [6]
نیازمندی‌ها	قاعده تولید شده دستی برای استخراج نمونه‌های آموزشی مناسب	استفاده از ویکی‌پدیا برای داده‌های آموزشی	تعداد نمونه اولیه برای آموزش و یک مدل اولیه	الگوهای نحوی	دیکشنری بزرگ از عبارات رابطه‌ای و الگوهایی برای محدودیت‌های نحوی	استفاده از ویکی‌پدیا برای داده‌های آموزشی	مجموعه کوچک لغوی مستقل از دامنه
چالش‌ها	نیاز به قواعد دستی، استخراج‌های بی‌معنی و استخراج‌های خالی از اطلاعات مفید	استخراج‌های بی‌معنی و استخراج‌های خالی از اطلاعات مفید	استخراج‌های بی‌معنی و استخراج‌های خالی از اطلاعات مفید	فقط کشف روابط واضح و مابین دو موجودیت	شناسایی نادرست آرگومان‌ها و وابستگی به گرامر زبان انگلیسی	وابسته به گرامر زبان، نیاز به مجموعه داده آموزشی	وابسته به گرامر زبان، استخراج موارد غیر حقیقی، آهسته‌تر از همه تکنیک‌های دیگر
ارزیابی	میانگین دقت ۷۵٪ ، خطای کمتر نسبت به KnowItAll و مقیاس‌پذیری بالا	با فراخوانی ۴۰٪، WOE^{parse}با دقت ۶۳٪ و WOE^{POS}با دقت ۵۵٪	فراخوانی بالا و دقت بالا با افزایش تعداد تکرار و اضافه شدن نمونه‌های اولیه	دقت ۸۸,۳٪ با فراخوانی ۴۵,۲٪	با فراخوانی ۴۰٪، دقت بالاتر از ۸۰٪ برای رابطه F1مربوط به Arg1، ۶۰٪ و Arg2، ۵۳٪ مربوط به صفحات وب	F1مربوطبه Arg1، ۸۱٪ و Arg2، ۷۲٪ مربوط به صفحات وب	دقت و فراخوانی بالایی نسبت به سیستم‌های دیگر
تکنیک‌های مورد استفاده	استفاده از سیستم RESOLVER	POSو Parser	استفاده از مدل‌های آماری، تکنیک‌های سطحی و ارزان‌تر برای تولید الگو	استفاده از CRF	محدودیت‌های نحوی و محدودیت‌های لغوی	ReVerb و ARGLEARNER	تجزیه وابستگی و برچسب گذاری معنایی کلمات

۲-۵- استخراج اطلاعات از متون فارسی

سیستم هدفمند مبتنی بر خروجی مرصاد دانشگاه شهید بهشتی [14]، برای استخراج اطلاعات از اخبار نظامی متون زبان فارسی استفاده می‌شود، ولی الگوهای استخراج حقایق در این سیستم بصورت دستی تعریف شده است. در این سیستم متون خبری در واحد پیش پردازش، توسط یک تجزیه‌گر نحوی زبان فارسی به اجزای فعل، فاعل، مفعول، گروه حرف اضافه‌ای، زمان و مکان تجزیه می‌شوند. بعد جملات با استفاده از یک شبکه واژگان معنایی مورد بررسی قرار می‌گیرند. سپس برای کلمات و عبارات مختلف، معادل‌های مفهومی آنها جایگزین می‌شود و ساختار مفهومی خبر ساخته می‌شود. این ساختار دقیقاً مشابه اطلاعات خام است و اطلاعی در این واحد از متون خبری کاسته نمی‌شود. واحد انطباق الگوها، جدول ساختار مفهومی را دریافت کرده و با توجه به محدودیت‌های الگوهای استخراج، جمله خبری را بررسی می‌کند، جملات فاقد اطلاع مفید حذف می‌شوند. برای سیستم مرصاد بیش از ۳۰ الگو طراحی شده است.در شکل (۲-۶) نمونه‌هایی از الگوهای سیستم مرصاد نمایش داده شده است.

Action	حمله کردن
Subj.	
Comp.	با ...
Comp.	به ...
Time	
Place	

الف

Action	کشته شدن
Subj.	
Comp.	بر ... در اثر ...
Time	
Place	

ب

Action	برعهده گرفتن
Subj.	
Obj.	مسئولیت
Time	
Place	

ج

شکل ۲-۶: نمونه‌هایی از الگوهای مرصاد [14]

جدول الگوها همانند ساختار مفهومی خبر شامل بخش‌های فعل، فاعل، مفعول، گروه حرف اضافه‌ای، زمان و مکان است. وجود فعل در یک الگو الزامی است و نشان دهنده یک رویداد می‌باشد.

در انطباق الگوها با ساختار مفهومی خبر، دو دسته انطباق دقیق و جزئی انجام می‌شود. در انطباق دقیق جمله باید با تمام محدودیت‌های الگو سازگار باشد، در غیر این صورت هیچ اطلاعی از آن جمله استخراج نمی‌شود. در انطباق جزئی برای همه اجزای جمله، انطباق دقیق لازم نیست.

در مقاله [42] روش‌هایی برای استخراج روابط معنایی میان فعل و وابسته‌های آن از متون زبان فارسی توضیح داده شده است. در این مقاله سعی در استخراج مجموعه نقش‌های معنایی با استفاده از سه روش مبتنی بر ریخت‌شناسی (مطالعه ساختار لغات) با استفاده از فرهنگ واژگانی زبان فارسی، مبتنی بر تعمیم با استفاده از پایگاه داده ساختار وابستگی افعال و مبتنی بر قاعده و تعمیم با استفاده از پیکره متون زبان فارسی را دارد. نقش‌های معنایی مانند عامل، همراه، حالت، مبدا و ...در نظر گرفته شده است. این نقش‌ها برای برچسب‌گذاری معنایی کلمات استفاده می‌شود. همچنین در مقاله [43] روش‌های مختلف استخراج بی‌ناظر ظرفیت فعل در زبان فارسی بررسی شده است و راه‌حل‌هایی برای

مسائلی در خصوص یافتن فعل در متون زبانی و ابهامات موجود در شناخت ظرفیت فعل در زبان فارسی پیشنهاد شده است.

در مقاله [44] با استفاده از هستان‌شناسی برای متون زبان فارسی حاشیه‌نویسی انجام می‌شود. در این مقاله با استفاده از موتورهای جستجو و دایرةالمعارف‌ها نمونه‌ها شناسایی می‌شوند و با استفاده از روش‌های آماری و مبتنی بر الگو حاشیه‌نویسی می‌شوند.

۲-۶- شبکه واژگان

در حال حاضر پروژه شبکه واژگان برای بیش از ۴۰ زبان منتشر شده است. به عنوان مثال پروژه‌ای با عنوان شبکه واژگان چندگانه[۴۰] [45]، زبان‌هایی همچون ایتالیایی، اسپانیایی، پرتغالی و رومانی را پشتیبانی می‌کند.شبکه واژگان رسمی فارسی، FarsNet[24] است، که در حال حاضر این پایگاه دانش حاوی ۳۴،۰۵۹ مجموعه مترادف است.

مرحله اول ایجاد شبکه واژگان برای هر زبانی، گردآوری لغات و مفاهیم آن شبکه واژگان است. مفاهیم و لغات را می‌توان به دو دسته تقسیم کرد [24].

- مفاهیم مشترک که در تمام زبان‌ها وجود دارند.
- مفاهیم مستقل زبان فارسی که تنها در بین فارسی‌زبانان رواج دارند.

مفاهیم بخش اول را می‌توان از شبکه واژگان‌های زبان‌های دیگر استخراج کرد، و مفاهیم زبان فارسی را نیز می‌توان از پیکرهای زبان و فرهنگ لغات استخراج نمود.

برای ایجاد شبکه واژگان زبان رومانیایی برای قسمتی از شبکه واژگان BalkaNet، در مرحله‌ی اول مفاهیم اساسی مشترک مابین زبان‌های موجود در شبکه واژگان چند زبانه در نظر گرفته شده

---

[۴۰] MultiWordNet

است و بدنبال آن مفاهیم مختص هر زبان را اضافه می‌شود؛ البته یکی از کاربردهای شبکه واژگان چند زبانه ترجمه ماشینی است که توسط آن بتوان ترجمه یک مفهومی را به درستی در یک متن بدست آید. برای این منظور، بایستی واژگان معنایی بسیار عظیمی در اختیار باشد و همچنین با توجه به اضافه شدن لغات جدید در گذر زمان به هر زبان باید واژگان معنایی، بروزرسانی شود و دائما گسترش یابد [46].

برای ایجاد شبکه واژگان ژاپنی از شبکه واژگان انگلیسی استفاده شده است.در ابتدا از مجموعه‌های هم‌خانواده هسته آغاز شده و مجموعه‌های هم‌خانواده معتبر در زبان ژاپنی استخراج شده است به این ترتیب از ۱۱۷،۶۵۹ مجموعه هم‌خانواده در شبکه واژگان نسخه سوم ۶۲،۸۳۲ مجموعه هم‌خانواده ژاپنی بدست آمده است [47].

در شبکه واژگان فارس‌نت برای ایجاد مجموعه‌های هم‌خانواده در دسته اسم، از ترجمه مجموعه‌های هم‌خانواده هم‌معنی انگلیسی استفاده می‌شود. اسم‌هایی که از ترجمه مجموعه‌های هم‌خانواده شبکه واژگان انگلیسی تولید نشده‌اند، توسط اسامی رایج از PLDB[۴۱] و پیکره‌ی بیجن‌خان اضافه می‌شوند. روابط بین مجموعه هم‌خانواده‌ها، از متن خام یا برچسب‌زده‌شده از دو منبع (مقالات ویکی‌پدیا و پیکره بیجن‌خان) استخراج می‌شوند. بخشی از شبکه واژگان فارس‌نت، استخراج نیمه-خودکار روابط مفهومی است که با استفاده از چهار رویکرد اساسی؛ مبتنی بر الگو، روش‌های ساختاری، روش آماری و روش مبتنی بر شباهت استخراج خودکار روابط انجام می‌شود [24].

در روش مبتنی بر الگو، بر اساس یک سری الگوها و کلمات کلیدی، روابط بدست می‌آید. در این روش دقت بالا است ولی فراخوانی آنها در یک پیکره محدود است. روش مبتنی بر ساختار، از قسمت‌های مفید متن مانند جداول و لینک‌ها استفاده می‌شود. در روش آماری برای استخراج روابطی

[۴۱] Persian Linguistic Database

که هم‌زمان اتفاق می‌افتند، استفاده می‌شود. برای استخراج روابط بین هر جفت از کلمات لازم است که در پیکره مانند بیجن‌خان جستجو شود تا مشخص شود که چند درصد از مواقع، این دو کلمه به‌صورت هم‌زمان استفاده‌شده‌است. در صورتیکه این تعداد از یک حد آستانه بیشتر باشد، این دو کلمه به عنوان هم‌رخداد تلقی می‌شوند [24].

در دسته صفت، صفت‌ها به صفت‌های ساده و صفت‌هایی که از اضافه کردن پسوند به کلمات دیگر مانند اسم بدست می‌آیند، تقسیم می‌شوند. برای مثال در فارس‌نت، صفت‌ها را به ۱۲ قسمت معنایی و ۵۷ زیرکلاس تقسیم نموده است؛ مانند صفات ادراکی، زمانی، حرکتی، ماده، احساسی و .... در دسته فعل، به نوع فعل از نظر ساده و مرکب بودن و لازم و متعدی بودن و ... توجه شده است[24].

در مقاله [32] کلمات فارسی از پیکره‌های متنی استخراج می‌شوند، مانند مقاله [24] کلمات مرتبط را با استفاده از هم‌زمانی رخداد کلمات شناسایی می‌کند. به این ترتیب گروه‌هایی از کلمات فارسی مشابه ایجاد می‌شود. برای این گروه‌ها، نزدیک‌ترین مجموعه هم‌خانواده در شبکه واژگان انگلیسی بر اساس کلمات مجموعه هم‌خانواده و توضیحات انتخاب می‌شود.

۲-۷- خلاصه فصل

در این فصل، پس از مفاهیم اولیه، مرور ادبیات در چهار بخش انجام شد. ابتدا روش‌های استخراج اطلاعات مورد مطالعه و بررسی قرار گرفت. در این بخش ارزیابی و مقایسه برخی از سیستم‌های استخراج اطلاعات با توجه به روش استخراج اطلاعات مورد استفاده ارائه شد. در ادامه‌ی کارهای گذشته، سیستم‌های استخراج آزاد اطلاعات مورد بحث و بررسی قرار گرفت. در این بخش، به مشکلات و چالش‌های موجود در این سیستم‌ها پرداخته شد.

در بخش سوم استخراج اطلاعات از متون فارسی با توجه به مشکلات موجود در پردازش متون فارسی بیان شد. سیستم استخراج اطلاعات مرصاد مورد بحث قرار گرفت. در راه‌کار پیشنهادی به

شبکه واژگان نیاز است. کیفیت شبکه واژگان بر روی حقایق استخراج شده در قالب معنایی RDF تاثیر می‌گذارد، در نتیجه در بخش پایانی کارهای گذشته به صورت مختصر به چگونگی ایجاد شبکه واژگان پرداخته شد.

فصل سوم :

روش پیشنهادی

فصل ۳- روش پیشنهادی

در این فصل، روش پیشنهادی پایان‌نامه برای استخراج حقایق از متون فارسی در قالب RDFبه تفصیل شرح داده می‌شود. قبل از بررسی جزئیات سیستم، ویژگی‌های کلی سیستم پیشنهادی در جدول (۳-۱) ارائه شده است.

جدول ۳-۱: ویژگی‌های کلی سیستم پیشنهادی

سیستم	سطح استخراج اطلاعات	روش شناسایی روابط	روش استخراج حقایق	قالب حقایق
سیستم پیشنهادی	سطح جمله	استخراج آزاد اطلاعات	پردازش زبان طبیعی	RDF

۳-۱- انگیزه

۳-۱-۱- انگیزه استفاده از استخراج آزاد در سطح جمله

همانطور که در مقدمه مطرح شد، در روش پیشنهادی ازروش استخراج آزاد اطلاعات در سطح جمله استفاده می‌شود. در روش استخراج آزاد اطلاعات همه روابط با معنی ممکن کشف می‌شود و به یک سری روابط خاص محدود نمی‌شود.

با توجه به اینکه کلمات برای توصیف معنی خاص در سطح جمله از یکدیگر مستقل نیستند، استخراج حقایق در سطح جمله انجام می‌پذیرد. روابط بین موجودیت‌ها با توجه به جملات متن کشف می‌شوند.

۳-۱-۲- انگیزه استفاده از روش پردازش زبان طبیعی

در روش مبتنی بر الگو، الگوهای متنی با استفاده از روش‌های دستی، یادگیری با ناظر و یادگیری خودناظر ایجاد می‌شوند. در روش دستی، عبارات منظم یا قواعد، توسط خود انسان تهیه

می‌شود. در نتیجه سیستم‌ها هزینه‌بر هستند و مقیاس‌پذیر و قابل حمل نیستند. در روش‌های یادگیری باناظر، مجموعه آموزشی نمونه‌های برچسب خورده در دامنه خاص، به سیستم داده می‌شود. سیستم‌های DIPRE [36] و Snowball [3, 36] از این روش استفاده می‌کنند. در روش یادگیری خودناظر، برچسب گذاری مثال‌های آموزشی با استفاده از یک مجموعه کوچک از الگوهای استخراج مستقل از دامنه، انجام می‌شود. در این روش، به جای استفاده از داده آموزشی برچسب خورده، مثال‌ها انتخاب و برچسب‌گذاری می‌شوند [12].

از آنجاییکه مجموعه داده آموزشی برای متون زبان فارسی وجود ندارد، نمی‌توان الگوها و قواعد را به صورت خودکار بدست آورد و از طرفی تولید الگوها و یا قواعد به صورت دستی نیز هزینه‌بر است. اگر از روش تولید الگو خودناظر مانند سیستم Know-ItAll[12] استفاده شود، برای تولید داده‌های آموزشی به مثال‌های برچسب خورده نیاز است.

در روش مبتنی بر هستان‌شناسی، استخراج حقایق بر اساس هستان‌شناسی انجام می‌شود، که دامنه روابط حقایق محدود به هستان‌شناسی موردنظر است. در حالیکه برای زبان فارسی هستان‌شناسی کاملی وجود ندارد که بتوان از آن در استخراج حقایق استفاده کرد. علاوه بر این، یکی از مواردی که در روش پیشنهادی مورد توجه بوده است، عدم وجود محدودیت به روابط دامنه خاص می‌باشد.

با توجه به چالش‌های موجود در روش‌های مبتنی بر الگو و هستان‌شناسی، در روش پیشنهادی از روش مبتنی بر پردازش زبان طبیعی استفاده می‌شود.

از طرفی ابزار برچسب‌گذاری معنایی SRL برای متون زبان فارسی با دقت مناسبی وجود ندارد و همچنین استفاده از این ابزار برچسب‌گذاری معنایی موجب کند شدن سیستم استخراج حقایق می‌شود. با توجه به اینکه در تجزیه وابستگی ارتباط نحوی کلمات با یکدیگر مشخص می‌شود و

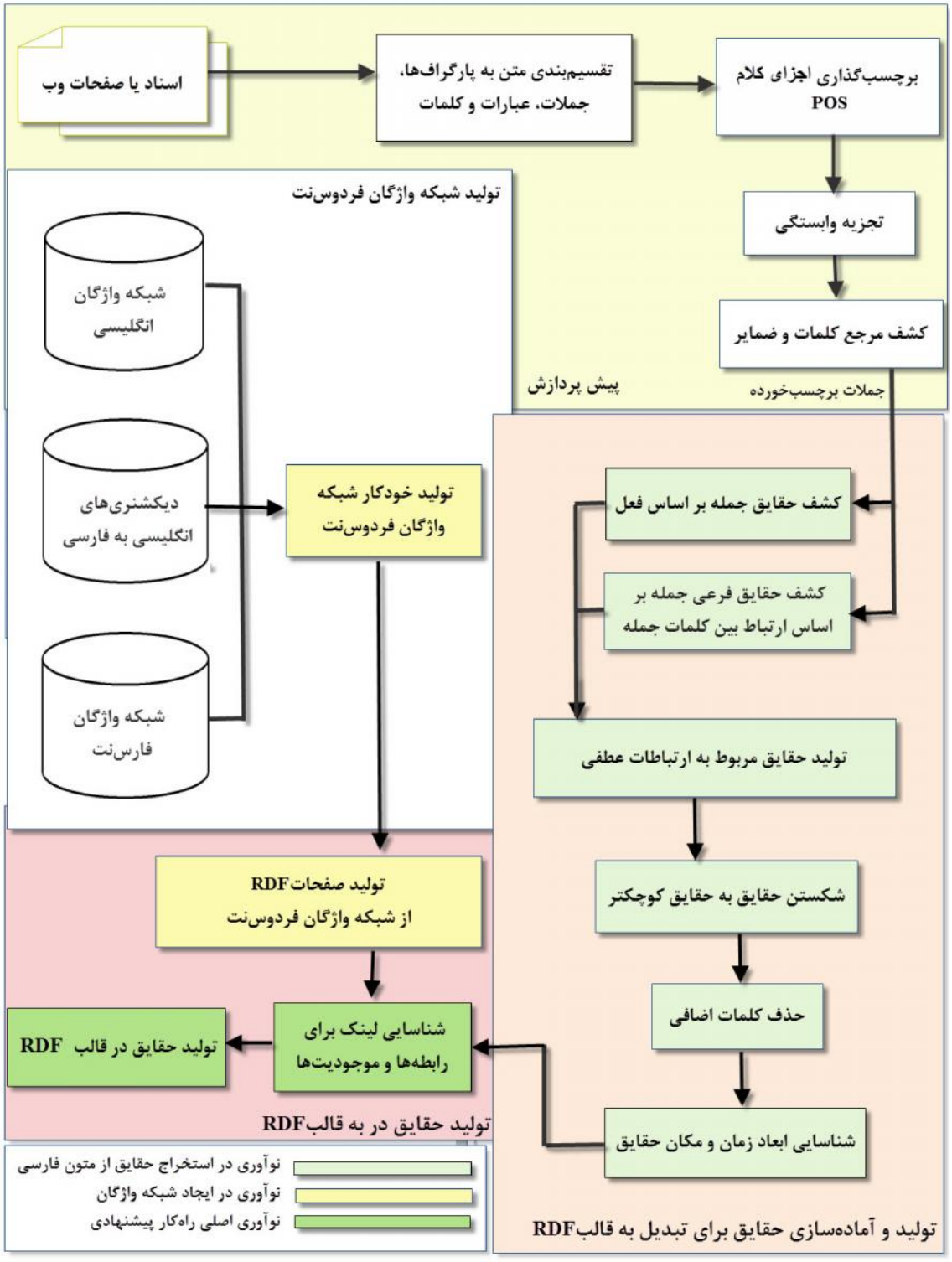
همچنین نقش‌هایی مانند فاعل و مفعول و تمیز و مسند و غیره برچسب‌گذاری می‌شوند. پس می‌توان با استفاده از تجزیه وابستگی اطلاعات مفید و روابط موجودیت‌ها را کشف نمود.

۳-۲- معماری سیستم پیشنهادی

معماری سیستم پیشنهادی در شکل (۳-۱) نشان داده شده است. مراحل اصلی در راه‌کار پیشنهادی عبارتند از:

- استخراج حقایق سه‌تایی از جملات
- آماده‌سازی حقایق برای تبدیل به قالب معنایی RDF
- ایجاد شبکه واژگان‌فردوس‌نت
- تولید صفحات RDF از شبکه واژگان
- شناسایی لینک برای حقایق

در ادامه، جزئیات هر کدام از مراحل راه‌کار پیشنهادی شرح داده می‌شود.



شکل ۳-۱ : معماری سیستم پیشنهادی

۳-۳- پیش‌پردازش

در بخش پیش‌پردازش، اسناد دریافتی به پاراگراف‌ها و جملات تقسیم‌بندی می‌شوند. همانطور که در بخش انگیزه بیان شد، استخراج حقایق سیستم پیشنهادی در سطح جمله می‌باشد، در نتیجه باید اطلاعات متن به جملات شکسته شود.

در این بخش، بعد از شناسایی کلمات هر جمله، برچسب‌گذاری اجزای کلام POS انجام می‌شود. برچسب‌گذاری اجزای کلام، عمل انتساب برچسب‌های نحوی به واژه‌ها و نشانه‌های تشکیل دهنده یک متن است. این برچسب‌ها نشان‌دهنده نقش کلمات و نشانه‌ها در جمله می‌باشند. درصد بالایی از واژگان از نقطه نظر برچسب‌گذاری نحوی دارای ابهام هستند، زیرا کلمات در جایگاه‌های مختلف برچسب‌های نحوی متفاوتی دارند. بنابراین برچسب‌گذاری نحوی، عمل ابهام‌زدایی از برچسب‌ها با توجه به متن مورد نظر است.

در قسمت تجزیه‌گر وابستگی، وابستگی‌های نحوی کلمات جمله مشخص می‌شود. تجزیه‌گرها با بهره‌گیری از دستورات گرامری زبان به تفکیک جملات متون به اجزای تشکیل‌دهنده‌ی آن، مشخص کردن نقش هر عبارت و لغت در متن و همچنین تشکیل درخت تجزیه برای جملات متن می‌پردازند. در تجزیه‌گر وابستگی علاوه بر نقش‌های کلمات، وابستگی‌های کلمات به لحاظ نقش آن‌ها در جمله نیز مشخص می‌شود.

اجزای هر جمله را می‌توان در قالب گروه‌های اسمی، فعلی، حرف اضافه‌ای و ... تقسیم‌بندی نمود. هر کدام از این گروه‌ها می‌توانند شامل زیرگروه دیگری می‌باشند. علاوه بر این،هر کدام نیز دارای روابطی می‌باشند، مثلا یک گروه اسمی می‌تواند متعلق به یک گروه فعلی باشد. در نتیجه‌ی این تقسیم‌بندی‌های سلسله مراتبی، می‌توان یک ساختار درخت‌گونه از جمله داشت که درخت تجزیه نام دارد. درخت تجزیه، درختی است که ساختار نحوی یک جمله را بر اساس برخی روابط گرامری موجود در آن به شکلی ساده و قابل فهم نمایش می‌دهد. ابزارهای مختلفی برای تجزیه جمله توسعه یافته‌اند

که خروجی اغلب آنها به صورت رشته‌ای شامل پرانتزهای تو در تو به همراه برچسب‌ها و کلمات می‌باشند. این مدل نمایش برای ورودی سیستم‌ها مناسب است، اما برای انسان خوانایی چندانی ندارد.

مرحله آخر در پیش‌پردازش، کشف مرجع ضمایر و کلمات مانند اختصارات است. مثلا ضمایر در جملات به سایر اسامی خاص متن اشاره می‌کنند. از آن‌جایی‌که استخراج اطلاعات در سطح جمله انجام می‌شود و اطلاعات هر حقیقت با توجه به کلمات هر جمله بدست می‌آید. در نتیجه وجود ضمایر در جملات موجب ابهام در حقایق می‌شود و از طرفی سیستم در بخش شناسایی لینک برای اطلاعات حقیقت در جهت تبدیل به قالب معنایی RDF با مشکل روبرو می‌شود. بنابراین در بخش پیش‌پردازش مرجع کلمات مانند ضمایر مشخص می‌شود.

جملات برچسب‌خورده در بخش پیش‌پردازش، مطابق پیکره تجزیه‌ی وابستگی دادگان [48] است. پیکره تجزیه‌ی وابستگی دادگان بر اساس قالب معیار همایش زبان‌شناسی رایانه‌ای و پردازش زبان طبیعی [49] بر روی پیکره‌های وابستگی فراهم شده است.

۳-۴- استخراج حقایق سه‌تایی از جملات

در این مرحله از استخراج حقایق، جملات (ساده یا مرکب) از دو منظر بررسی می‌شوند.

- حقایق اصلی جمله که بر اساس افعال جمله شناسایی می‌شوند.
- حقایق فرعی که بر اساس ارتباط بین کلمات جمله کشف می‌شوند.

۳-۴-۱- استخراج حقایق اصلی جمله

در تجزیه‌ی وابستگی ارتباط نحوی بین کلمات جمله مشخص می‌شود. کلمه اصلی جمله در تجزیه‌ی وابستگی به عنوان ریشه برچسب‌گذاری می‌شود. معمولا ریشه جمله، فعل اصلی جمله است. همانگونه که در مثال زیر مشاهده می‌شود، در ابتدا سیستم پیشنهادی، با توجه به برچسب‌های تجزیه وابستگی، همه کلماتی که با ریشه ارتباط مستقیم دارند، را مشخص و بررسی می‌کند.

مثال: جمله «سال نو آمد و کنار سفره هفت‌سین نشست.» برچسب‌گذاری شده است.

1	سال	سال	N	IANM	attachment=ISO|number=SING|senID=23944	3	SBJ	_	_
2	نو	نو	ADJ	AJP	attachment=ISO|senID=23944	1	NPOSTMOD	_	_
3	آمد	آمد#آ	V	ACT	person=3|attachment=ISO|number=SING|tma=GS|senID=23944	4	PREDEP	_	_
4	و	و	CONJ	CONJ	attachment=ISO|senID=23944	8	VCONJ	_	_
5	کنار	کنار	PREP	PREP	attachment=ISO|senID=23944	8	ADV	_	_
6	سفرهٔ	سفره	N	IANM	attachment=ISO|number=SING|senID=23944	5	POSDEP	_	_
7	هفت‌سین	هفت‌سین	N	IANM	attachment=ISO|number=SING|senID=23944	6	MOZ	_	_
8	نشست	نشست#تشین	V	ACT	person=3|attachment=ISO|number=SING|tma=GS|senID=23944	0	ROOT	_	_
9	.	.	PUNC	PUNC	attachment=ISO|senID=23944	8	PUNC	_	_

سیستم پیشنهادی با استفاده از ارتباط بین کلمات جمله، اجزای حقایق را مشخص می‌کند.فاعل جمله بعنوان نهاد حقیقت و فعل به عنوان مسند حقیقت انتخاب می‌شود.

اگر فاعل بین کلماتی که با فعل ارتباط مستقیم دارند، وجود نداشت، حروف ربطی که به فعل جمله متصل شده‌اند، بررسی می‌شوند. اگر حرف ربط دارای نقش «هم‌پایه‌ی فعل» باشد، فعلی که ارتباط وابستگی به حرف ربط مورد نظر داشته باشد، مورد بررسی قرار می‌گیرد. مانند فعل «آمد».فاعل این فعل به عنوان فاعل فعل ریشه هم انتخاب می‌شود. اگر این فعل دارای فاعل نبود، همین عملیات برای فعل مورد نظر «آمد» انجام می‌شود.

گزاره هر حقیقت با توجه به نوع وابستگی نحوی کلمات مشخص می‌شود. اگر برای حقیقتی گزاره شناسایی نشود، گزاره آن تهی است. گزاره‌ها بر اساس نقش نحوی با توجه به نظر اساتید زبان‌شناسی به صورت زیر تقسیم‌بندی شده‌اند:

- قید: اگر کلمه‌ای با نقش نحوی قید به فعل متصل باشد، یک حقیقت با نوع قیدی ایجاد می‌شود. گزاره این حقیقت، قید جمله می‌باشد.
    - مثال: «دیروز علی رفت.»
        - («علی»، «رفت»، «دیروز»)
- مسند: اگر کلمه‌ای با نقش نحوی مسند به فعل متصل باشد، یک حقیقت با نوع مسندی ایجاد می‌شود. گزاره این حقیقت، مسند جمله می‌باشد.
    - مثال: «هوا سرد است.»
        - («هوا»، «است»، «سرد»)
- مفعول: اگر کلمه‌ای با نقش نحوی مفعول به فعل متصل باشد، یک حقیقت با نوع مفعولی ایجاد می‌شود. گزاره این حقیقت، مفعول جمله می‌باشد.اگر مفعول، مفعول دوم باشد، علاوه بر تولید حقیقت مربوط به مفعول دوم، یک حقیقت با گزاره «مفعول-مفعول» هم ایجاد می‌شود تا اطلاعات مفید حذف نشود.
    - مثال: «احمد سیب را خورد.»
        - («احمد»، «خورد»، «سیب را»)
    - مثال: «سودابه کتاب را به شیما هدیه داد.»
        - («سودابه»، «هدیه داد»، «کتاب را»)
        - («سودابه»، «هدیه داد»، «به شیما»)
        - («سودابه»، «هدیه داد»، «کتاب را به شیما»)
- تمیز: اگر کلمه‌ای با نقش نحوی تمیز به فعل متصل باشد، یک حقیقت با عناصر اصلی جمله که به فعل متصل هستند مانند مفعول، مسند همراه با تمیز ایجاد می‌شود.

- مثال: «فرماندهان از این منطقه به عنوان پناهگاه نام می‌برند.»
  - («فرماندهان»، «نام می‌برند »، «از این منطقه»)
  - («فرماندهان»، «نام می‌برند »، «از این منطقه به عنوان پناهگاه »)

1	فرماندهانفرمانده	N	ANM	attachment=ISO|number=PLUR|senID=23484	9	SBJ	_	_
2	ازاز	PREP	PREP	attachment=ISO|senID=23484	9	VPP	_	_
3	این این	PREM	DEMAJ	attachment=ISO|senID=23484	4	NPREMOD	_	_
4	منطقه منطقه	N	IANM	attachment=ISO|number=SING|senID=23484	2	POSDEP	_	_
5	بهبه	PREP	PREP	attachment=ISO|senID=23484	9	TAM	_	_
6	عنوان عنوان	N	IANM	attachment=ISO|number=SING|senID=23484	5	POSDEP	_	_
7	پناهگاه پناهگاه	N	IANM	attachment=ISO|number=SING|senID=23484	6	MOZ	_	_
8	نام نام	N	IANM	attachment=ISO|number=SING|senID=23484	9	NVE	_	_
9	برد#بر می‌برند	V	ACT	person=3|attachment=ISO|number=PLUR|tma=H|senID=234840	ROOT	_	_	

- بند افزوده‌ی فعل: اگر کلمه‌ای با نقش نحوی بند افزوده‌ی فعل به فعل متصل شود، یک حقیقت با عناصر اصلی جمله همراه با بند افزوده‌ی فعل تولید می‌شود.
  - مثال: «اگر مینا بیاید، نرگس خوشحال می‌شود»
    - («نرگس»، «می‌شود»، «خوشحال»)
    - («نرگس»، «می‌شود»، « اگر مینا بیاید، خوشحال»)
    - («مینا»، «بیاید»، «»)

اگر بعضی نقش‌های نحوی مانند قید، به تنهایی در قسمت گزاره حقیقت قرار بگیرند، حقایقی بی‌معنی تولید می‌شود، در بخش آماده‌سازی حقایق برای تبدیل به قالب RDF، قیدهای زمان و مکان

در جمله شناسایی می‌شوند و ابعاد زمان و مکان حقایق مشخص می‌شود. در نتیجه چالش‌هایی مانند حقایق بی‌معنی مربوط به قیدهای مکان و زمان رفع می‌شود.

البته کلماتی مانند «آیا، بنابراین، سپس و ...» برای گزاره حقیقت در نظر گرفته نمی‌شوند.

1	همچنینهمچنین	ADV	SADV	attachment=ISO|senID=24024	8	ADV	_	_	
3	سپس سپس	ADV	SADV	attachment=ISO|senID=24041	15	ADV	_	_	
11	بلکه بلکه	ADV	SADV	attachment=ISO|senID=24058	22	ADV	_	_	
19	متأسفانهمتأسفانه	ADV	SADV	attachment=ISO|senID=24066	37	ADV	_	_	
24	خوشبختانهخوشبختانه	ADV	SADV	attachment=ISO|senID=24071	29	ADV	_	_	
1	البته البته	ADV	SADV	attachment=ISO|senID=24112	19	ADV	_	_	
1	آیا آیا	PART	PART	attachment=ISO|senID=43362	16	PART	_	_	

مثال: جمله‌ی «رئیس بیمه مرکزی جمهوری اسلامی ایران خاطرنشان کرد: بر اساس آخرین آمار استخراج شده ضریب نفوذ بیمه ۱ درصد است.» برچسب‌گذاری شده است. سیستم پیشنهادی حقایق اصلی زیر را استخراج می‌کند:

- («رئیس بیمه مرکزی جمهوری اسلامی ایران »، «خاطرنشان کرد »، «بر اساس آخرین آمار استخراج شده ضریب نفوذ بیمه ۱ درصد است»)
- («ضریب نفوذ بیمه »، «است»، «۱ درصد »)
- («ضریب نفوذ بیمه »، «است»، «بر اساس آخرین آمار استخراج شده »)

1	رئیسرئیس _	N	ANM	attachment=ISO|number=SING|senID=23624	8	SBJ	_	
2	بیمه بیمهٔ _	N	IANM	attachment=ISO|number=SING|senID=23624	1	MOZ	_	
3	مرکزی مرکزی	ADJ	AJP	attachment=ISO|senID=23624	2	NPOSTMOD	_	_
4	جمهوری جمهوری MOZ	N	IANM	attachment=ISO|number=SING|senID=23624	2	_	_	

5	اسلامی	اسلامی	ADJ	AJP	attachment=ISO|senID=23624	4	NPOSTMOD	_	_
6	ایران	ایران	N	IANM	attachment=ISO|number=SING|senID=23624	4	MOZ	_	_
7	خاطرنشان	خاطرنشان	N	IANM	attachment=ISO|number=SING|senID=23624	8	NVE	_	_
8	کرد	کرد#کن	V	ACT	person=3|attachment=ISO|number=SING|tma=GS|senID=23624	0	ROOT	_	_
9	:	:	PUNC	PUNC	attachment=ISO|senID=23624	8	PUNC	_	_
10	بر	بر	PREP	PREP	attachment=ISO|senID=23624	20	ADV	_	_
11	اساس	اساس	N	IANM	attachment=ISO|number=SING|senID=23624	10	POSDEP	_	_
12	آخرین	آخرین	PRENUM	PRENUM	attachment=ISO|senID=23624	13	NPREMOD	_	_
13	آمار	آمار	N	IANM	attachment=ISO|number=SING|senID=23624	11	MOZ	_	_
14	استخراج‌شده	استخراج‌شده	ADJ	AJP	attachment=ISO|senID=23624	13	NPOSTMOD	_	_
15	ضریب	ضریب	N	IANM	attachment=ISO|number=SING|senID=23624	23	SBJ	_	_
16	نفوذ	نفوذ	N	IANM	attachment=ISO|number=SING|senID=23624	15	MOZ	_	_
17	بیمه	بیمه	N	IANM	attachment=ISO|number=SING|senID=23624	16	MOZ	_	_
18	1	1	PRENUM	PRENUM	attachment=ISO|senID=23624	19	NPREMOD	_	_
19	درصد	درصد	N	IANM	attachment=ISO|number=SING|senID=23624	20	MOS	_	_
20	است	است#	V	ACT	person=1|attachment=ISO|number=SING|tma=H|senID=23624	8	VCL	_	_
21	.	.	PUNC	PUNC	attachment=ISO|senID=23624	8	PUNC	_	_

۳-۴-۲- استخراج حقایق فرعی جمله

در این بخش حقایق از گروه‌های اسمی استخراج می‌شود. گروه‌های اسمی شامل اطلاعات مربوط به بدل، ترکیبات مضاف و مضاف‌الیه و موصوف و صفت می‌باشند. به علت اینکه برای شناسایی لینک برای نهاد و گزاره حقایق به ترکیبات مضاف و مضاف‌الیه نیاز است و مضاف‌الیه از اطلاعات

حقایق اصلی حذف نمی‌شود، سیستم پیشنهادی اطلاعات مربوط به ترکیبات موصوف و صفت و ترکیبات مضاف و مضاف‌الیه را با استفاده از تجزیه وابستگی استخراج می‌کند. البته موصوف و مضاف باید اسم باشند.

حقایق ترکیبات موصوف و صفت: اگر در درخت تجزیه وابستگی برای صفت برچسب وصفی تعیین شده باشد، یک حقیقت جدید ایجاد می‌شود. نهاد این حقیقت، موصوف و گزاره آن صفت است. مسند حقیقت با توجه به ترکیب موصوف و صفت، فعل «است» انتخاب می‌شود. در این سیستم اگر حقیقت موصوف و صفت به حقایق اضافه شود، به درخواست کاربر، صفت از حقایق اصلی در بخش قبلی حذف می‌شود و در آنجا فقط موصوف به تنهایی قرار می‌گیرد.

- مثال:«ایران زیبا»
- («ایران»، «است»، «زیبا»)

حقایق ترکیبات مضاف و مضاف‌الیه: در این بخش مضاف‌الیه در قسمت نهاد حقیقت و مضاف بعنوان گزاره حقیقت قرار می‌گیرند. در قسمت مسند حقیقت، فعل «دارد» گذاشته می‌شود.

- مثال: «رئیس مجلس»
- («مجلس»،«دارد»، «رئیس»)

در جمله‌ی «رئیس بیمه مرکزی جمهوری اسلامی ایران خاطرنشان کرد: بر اساس آخرین آمار استخراج شده ضریب نفوذ بیمه ۱ درصد است.» حقایق فرعی زیر استخراج می‌شود:

- («بیمه مرکزی جمهوری اسلامی ایران»،«دارد»، «رئیس»)
- («جمهوری اسلامی ایران»،«دارد»، «بیمه مرکزی»)
- («بیمه»،«است»، «مرکزی»)
- («جمهوری»،«است»، «اسلامی»)

- («ایران»،«دارد»، «جمهوری اسلامی »)
- («آمار استخراج شده»،«دارد»، «اساس»)
- («آمار»،«است»، «استخراج شده»)
- («نفوذ بیمه »،«دارد»، «ضریب»)
- («بیمه»،«دارد»، «نفوذ»)

استخراج حقایق با استفاده از بدل: بدل‌ها، عبارت‌های اسمی هستند که اغلب در ادامه‌ییک اسم و یا گروه اسمی آمده است و نقش توضیحی و مکملی را برای آن گروه اسمی دارد.

مثال: عبارت «نوشته امجدیان» در جمله‌ی «"مسافر تنها" نوشته امجدیان روایتی داستانی از زندگی شهید آیت‌الله اصفهانی است.»، بدل است. حقایق زیر از جمله مذکور شناسایی می‌شوند:

- («مسافر تنها »،«است»، «روایتی داستانی از زندگی شهید آیت‌الله اصفهانی »)
- («مسافر تنها»،«است»، «نوشته امجدیان»)
- («مسافر تنها»،«است»، «روایتی داستانی»)
- («مسافر»،«است»، «تنها»)
- («امجدیان»،«دارد»، «نوشته»)
- («روایتی»،«است»، «داستانی»)
- («اصفهانی»،«است»، «شهید»)
- («اصفهانی»،«است»، «آیت‌الله»)
- («اصفهانی»،«دارد»، «زندگی»)

1	“	“	PUNC	PUNC	attachment=ISO|senID=24187	2	PUNC	_	_
2	مسافر	مسافر	N	ANM	attachment=ISO|number=SING|senID=24187	14	SBJ	_	_
3	تنها	تنها	ADJ	AJP	attachment=ISO|senID=24187	2	NPOSTMOD	_	_

4	"	"	PUNC	PUNC	attachment=ISO|senID=24187	2	PUNC	_	_
5	نوشتهٔ	نوشته	N	IANM	attachment=ISO|number=SING|senID=24187	2	APP	_	_
6	امجدیان	امجدیان	N	ANM	attachment=ISO|number=SING|senID=24187	5	MOZ	_	_
7	روایتی	روایت	N	IANM	attachment=ISO|number=SING|senID=24187	14	MOS	_	_
8	داستانی	داستانی	ADJ	AJP	attachment=ISO|senID=24187	7	NPOSTMOD	_	_
9	از	از	PREP	PREP	attachment=ISO|senID=24187	7	NPP	_	_
10	زندگی	زندگی	N	IANM	attachment=ISO|number=SING|senID=24187	9	POSDEP	_	_
11	شهید	شهید	IDEN	IDEN	attachment=ISO|number=SING|senID=24187	13	PREDEP	_	_
12	آیت‌الله	آیت‌الله	IDEN	IDEN	attachment=ISO|number=SING|senID=24187	13	PREDEP	_	_
13	اصفهانی	اصفهانی	N	ANM	attachment=ISO|number=SING|senID=24187	10	MOZ	_	_
14	است	#است	V	ACT	person=3|attachment=ISO|number=SING|tma=H|senID=24187	0	ROOT	_	_

۳–۴–۳– چالش‌ها

چالش‌های سیستم پیشنهادی در استخراج حقایق اصلی عبارتند از:

- شناسایی نهاد حقیقت: هنگامی که جملات پیچیده باشند و نهاد فعل مشخص نباشد.
- شناسایی گزاره حقیقت: هنگامی که گزاره یک فعل از نظر معنایی در قسمت یک فعل دیگر باشد، ولی ارتباط وابستگی با فعل مذکور نداشته باشد.

۳–۵– آماده‌سازی حقایق برای تبدیل به قالب معنایی RDF

با توجه به این موضوع که این حقایق برای مدل داده‌ای RDF آماده‌سازی می‌شوند، در نتیجه باید تا حد امکان از کلمات و یا حروف اضافی خودداری شود؛ همچنین قسمت‌های نهاد، مسند و گزاره

حقایق شکسته و کوچک شوند. در نتیجه حروف ربطی در قسمت‌های مختلف یک حقیقت مورد بررسی قرار می‌گیرند. بنابراین حقایق جدید با شکسته شدن قسمت‌های نهاد، مسند و گزاره حقایق بدست می‌آید. این قسمت‌های حقایق طوری شکسته می‌شوند که تغییری در قسمت‌های دیگر حقیقت نیاز نباشد.

در جمله «واحدهای صنعتی موفق و شرکت‌های خدماتی سبز در تهران انتخاب و معرفی شدند.» حقیقت اصلی زیر استخراج می‌شود:

- («واحدهای صنعتی موفق و شرکت‌های خدماتی سبز»،«انتخاب و معرفی شدند»، «در تهران »)

در این حقیقت قسمت‌های نهاد و مسند حقیقت دارای حروف ربطی هستند، در نتیجه این حقیقت می‌تواند به حقایق کوچکتری تبدیل شود.

- («واحدهای صنعتی موفق»،«انتخاب شدند»، «در تهران »)
- («شرکت‌های خدماتی سبز»،«انتخاب شدند»، «در تهران »)
- («واحدهای صنعتی موفق»،«معرفی شدند»، «در تهران »)
- («شرکت‌های خدماتی سبز»،«معرفی شدند»، «در تهران »)

1	واحدهای	واحد	N	IANM	attachment=ISO|number=PLUR|senID=35027		13	SBJ	
	_	_							
2	صنعتی	صنعتی	ADJ	AJP	attachment=ISO|senID=35027	1	NPOSTMOD	_	_
3	موفق	موفق	ADJ	AJP	attachment=ISO|senID=35027	1	NPOSTMOD	_	_
4	و	و	CONJ	CONJ	attachment=ISO|senID=35027	1	NCONJ	_	_
5	شرکت‌های	شرکت	N	IANM	attachment=ISO|number=PLUR|senID=35027		4		
	POSDEP	_	_						
6	خدماتی	خدماتی	ADJ	AJP	attachment=ISO|senID=35027	5	NPOSTMOD	_	_
7	سبز	سبز	ADJ	AJP	attachment=ISO|senID=35027	5	NPOSTMOD	_	_
8	در	در	PREP	PREP	attachment=ISO|senID=35027	13	ADV	_	_

9	تهران	تهران	N	IANM	attachment=ISO|number=SING|senID=35027	8	POSDEP	_	_
10	انتخاب	انتخاب	N	IANM	attachment=ISO|number=SING|senID=35027	13	NVE	_	_
11	و	و	CONJ	CONJ	attachment=ISO|senID=35027	10	NCONJ	_	_
12	معرفی	معرفی	N	IANM	attachment=ISO|number=SING|senID=35027	11	POSDEP	_	_
13	شدند	کرد#کن	V	PASS	person=3|attachment=ISO|number=PLUR|tma=GS|senID=35027	0	ROOT	_	_
14	.	.	PUNC	PUNC	attachment=ISO|senID=35027	13	PUNC	_	_

برای تبدیل حقایق به مدل داده‌ای RDF، باید شناسه یکتای منبع(URI) مناسب برای قسمت‌های نهاد، مسند و گزاره شناسایی شود. قسمت گزاره یک حقیقت می‌تواند مقدار ثابت باشد. اگر لینکی برای گزاره شناسایی نشود، گزاره به صورت یک رشته بدون پیوند ذخیره می‌شود.

اگر یک کلمه از بخش‌های مختلف حقیقت حذف شود، ممکن است، معنی جمله تغییر کند. برای مثال جمله «شرکت الف در تهران است» دارای حقیقت زیر می‌باشد:

- («شرکت الف»، «است»،«در تهران»)

حال اگر به موجودیت «تهران» لینک داده شود، معنی جمله تغییر می‌کند، زیرا حرف «در» نادیده گرفته شده است. در چنین مواقعی اگر گزاره حقیقت قید باشد و یا دارای حروف اضافه مانند «در» و «از» و ... باشد، باید نوع کلمه از نظر معنایی زمان یا مکان مشخص شود. در ادامه بجای گزاره‌ی حقیقت یک عدد ثابت قرار می‌گیرد و همین عدد ثابت در نهاد حقایق مرتبط با حقیقت اصلی نشان دهنده ارتباط این حقایق با یکدیگر می‌باشد. مسند این حقایق با توجه به نقش معنایی کلمه مشخص می‌شود.

- («شرکت الف»، «است»،۱۲۳۴۵)
- (۱۲۳۴۵، مکان،«تهران»)

۳-۵-۱- چالش‌ها

در بخش شکستن حقایق به حقایق کوچکتر ممکن است، مشکلاتی بوجود بیاید:

- اگر کلماتی که با حروف ربطی به یکدیگر متصل شده باشند، دارای کلماتی با رابطه معنایی مشترکی باشند. مثلا دارای یک صفت مشترک باشند. ولی در تجزیه وابستگی صفت فقط با یکی از کلمات رابطه وابستگی دارد. در نتیجه در هنگام شکستن حقایق ممکن است بعضی اطلاعات از بعضی حقایق حذف شود. مثال: در جمله «واحدهای صنعتی شرکت‌های خدماتی موفق در تهران انتخاب و معرفی شدند.» حقیقت اصلی زیر استخراج می‌شود:
  - («واحدهای صنعتی و شرکت‌های خدماتی موفق»،«انتخاب و معرفی شدند»، «در تهران »)
  - صفت «موفق» در تجزیه وابستگی فقط به کلمه «واحدهای» برمی‌گردد،در نتیجه هنگامی که این حقیقت شکسته می‌شود، حقایق زیر تولید می‌شود:
    - («واحدهای صنعتی موفق»،«انتخاب شدند»، «در تهران »)
    - («شرکت‌های خدماتی»،«انتخاب شدند»، «در تهران »)
    - («واحدهای صنعتی موفق»،«معرفی شدند»، «در تهران »)
    - («شرکت‌های خدماتی»،«معرفی شدند»، «در تهران »)
- در بعضی موارد اگر کلماتی که با حروف ربطی به یکدیگر متصل شده‌اند، جدا شوند، معنی جمله تغییر می‌کند:
  - «آرش و بهاره با یکدیگر ازدواج کردند.»

- در بعضی جملات اگر ابعاد زمان و مکان شناسایی شود و حرف اضافه حذف شود، ممکن است در معنای جمله ابهام ایجاد شود. مثلا در جمله «کوروش از مشهد رفت» حقیقت زیر استخراج می‌شود:
  - («کوروش»، «رفت»،«از مشهد»)
  - اگر بعد مکان حقیقت شناسایی شود، حقایق زیر استخراج می‌شوند:
    - («کوروش»، «رفت»،۱۲۳۴۶)
    - (۱۲۳۴۶، مکان،«مشهد»)

۳-۶- ایجاد شبکه واژگان‌فردوس‌نت

روش پیشنهادی ایجاد شبکه واژگان فردوس‌نت بر مبنای استفاده از شبکه واژگان انگلیسی می‌باشد. با توجه به اینکه شبکه واژگان انگلیسی با صرف ساعت‌ها زمان و توسط انسان تولید شده است، بسیار دقیق و قابل اعتماد بوده و می‌تواند مبنای خوبی برای ساخت سایر شبکه واژگان قرار گیرد. ایده اصلی راه‌کار پیشنهادی در ایجاد شبکه فردوس‌نت، بر این محور استوار است که مفاهیم و موجودات پیرامون ما دربین زبان‌های مختلف یکسان می‌باشند. بعنوان مثال هر «Lion» یک «Animal» می باشد، اینکه هر شیر یک حیوان می‌باشد، یک مفهوم مستقل از زبان است.

بنابراین، می‌توان گفت تمامی مفاهیمی که در زبان انگلیسی وجود داشته (که در شبکه واژگان انگلیسی هم تعریف شده است)، در زبان فارسی هم موجود می‌باشد. به عبارت دیگر، مجموعه‌های هم‌خانواده که در حقیقت همان مفاهیم می‌باشند، در زبان فارسی هم به همان شکل باید وجود داشته باشند. کافی است در مثال ذکر شده برای کلمه «Lion» ترجمه «شیر» و برای «Animal» ترجمه «حیوان» انتخاب گردد. در این صورت رابطه‌ی IS-A ایی که در شبکه واژگان انگلیسی می‌باشد، در زبان فارسی هم برقرار خواهد بود.

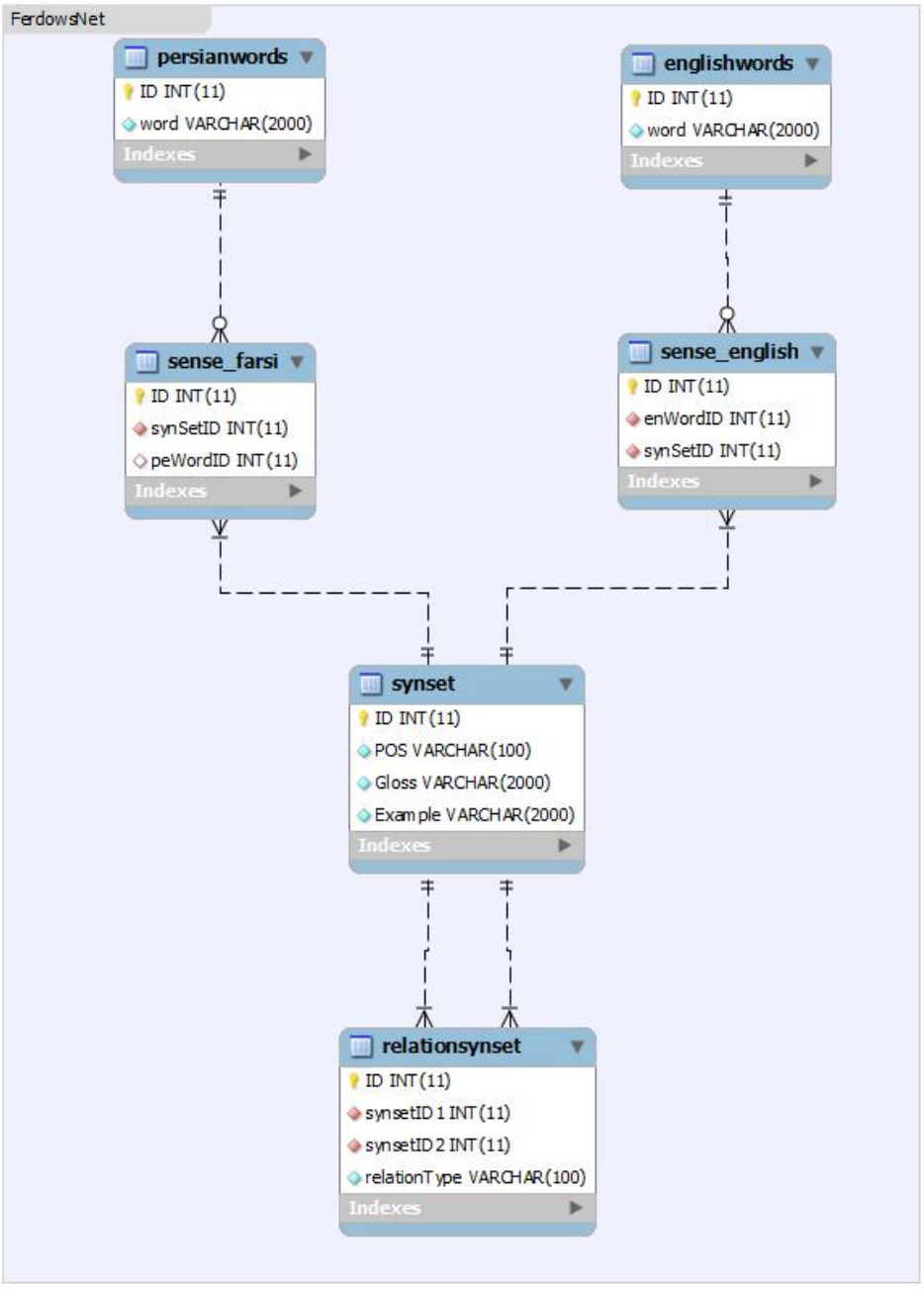
بنابراین درصورتی که برای هر کلمه انگلیسی، ترجمه مناسب آن درزبان فارسی برای مجموعه هم‌خانواده، پیدا شود، ضمن اینکه مجموعه‌های هم‌خانواده فارسی تا حد زیادی تولید می‌شود، می‌توان از روابطی که بین این مجموعه‌های هم‌خانواده در زبان انگلیسی تعریف شده است، نیز استفاده نمود. چراکه این روابط در زبان فارسی هم برقرار می‌باشند و درحقیقت مفاهیم و موجودیت‌ها و روابط بین آنها مستقل از زبان هستند. درجدول (۳-۲) نمونه‌ای از روابط بین مجموعه‌های هم‌خانواده در شبکه واژگان انگلیسی نشان داده شده است.

جدول ۳-۲: نمونه‌ای از روابط بین مجموعه‌های هم‌خانواده در شبکه واژگان انگلیسی

نقش کلمه	نام رابطه	توضیحات و مثال مربوطه
اسم	Hypernyms	Hypernyms ، y کلمه x است در صورتی که هر x نوعی از y باشد. (canine is a hypernym of dog)
	Hyponyms	Hyponyms ، y کلمه x است در صورتی که هر y نوعی از x باشد. (dog is a hypernym of canine)
	Holonym	Holonym ، y کلمه x است در صورتی که x بخشی از y باشد. (building is a holonym of window)
	Meronym	Meronym ، y کلمه x است در صورتی که y بخشی از x باشد. (window is a holonym of building)
	Coordinate terms	Coordinate term ،y کلمه x است در صورتی که xوy یک hypernym مشترک داشته باشند. (wolf is a coordinate term of dog)
فعل	Hypernyms	Hypernym ، y فعل x است در صورتی که فعالیت x نوعی از y باشد. (to perceive is an hypernym of to listen)
	Troponym	Troponym ، y فعل x است در صورتی که فعالیت y عمل x را به روشی انجام دهد. (to lisp is a Troponym of to talk)
	Entailment	Entailment ، y فعل x است در صورتی که با انجام x وابسته به y باشد (to sleep is entailed by to snore)
	Coordinate terms	Coordinate term ،y فعل x است در صورتی که xوy یک hypernym مشترک داشته باشند. (to lisp and to yell)

همانگونه در معماری سیستم پیشنهادی نشان داده شده است، برای ایجاد شبکه واژگان فردوس‌نت از شبکه واژگان انگلیسی، شبکه واژگان فارس‌نت، فرهنگ لغات انگلیسی به فارسی و فرهنگ واژگان مترادف و متضاد استفاده شده است.

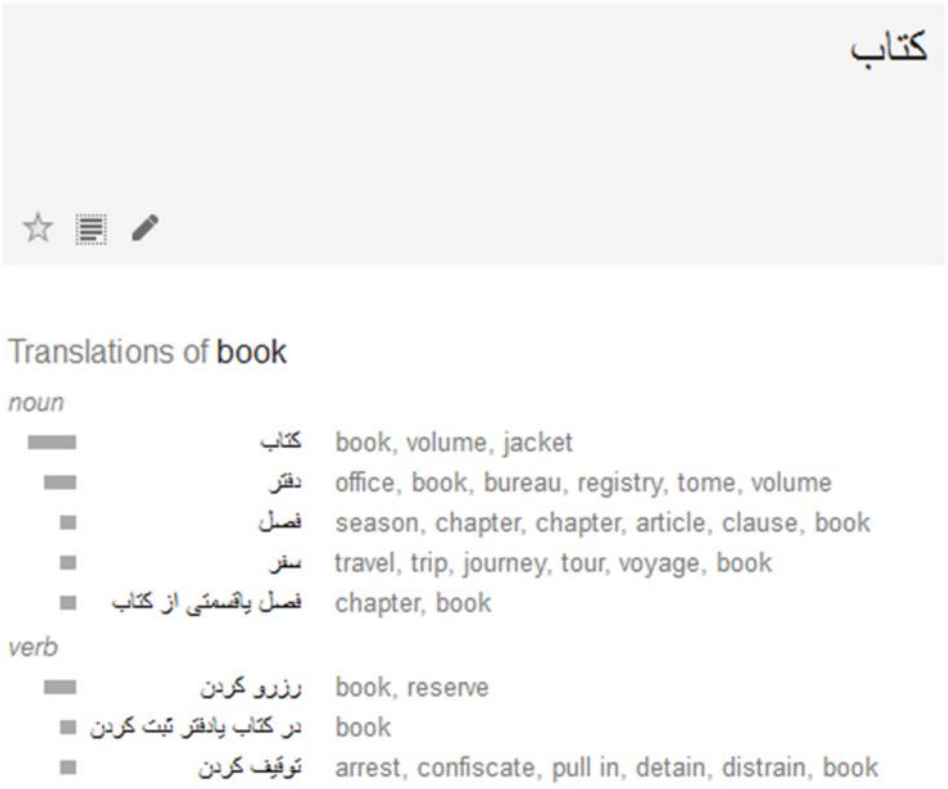
پایگاه داده شبکه فردوس‌نت در شکل (۳-۲) نمایش داده شده است. جدول روابط مجموعه‌های هم‌خانواده، روابط بین مجموعه‌های هم‌خانواده در دسته‌های مختلف اسم و فعل و صفت و قید نگهداری می‌کند، روابط بین مجموعه‌های هم‌خانواده فارسی همان روابط مجموعه‌های هم‌خانواده در شبکه واژگان انگلیسی می‌باشد.



شکل ۳-۲ : پایگاه داده شبکه فردوس‌نت

مجموعه‌های هم‌خانواده در شبکه واژگان فارس‌نت که با مجموعه‌های هم‌خانواده شبکه واژگان انگلیسی متناظر هستند، در شبکه واژگان فردوس‌نت استفاده می‌شوند. البته تعداد این مجموعه‌های هم‌خانواده نسبت به تعداد کل مجموعه‌های هم‌خانواده شبکه واژگان انگلیسی کم است.

برای ایجاد شبکه واژگان فردوس‌نت از چندین فرهنگ لغات انگلیسی به فارسی استفاده شده است. تعدادی از این فرهنگ لغات مانند TranslateGoogle برخط هستند. همه کلمات انگلیسی در شبکه واژگان انگلیسی با توجه به نقش آنها (فعل، اسم، صفت و قید) توسط این فرهنگ لغات برخط ترجمه شده‌اند. در شکل (۳-۳) ترجمه‌های مختلف کلمه «book» درTranslateGoogle نشان شده است.



شکل ۳-۳: یک نمونه از ترجمه‌ی کلمات در TranslateGoogle

کلمات در فرهنگ لغات با توجه به هر فرهنگ لغت رتبه‌بندی شده‌اند. مثلا ترجمه «کتاب» برای کلمه «book» در نقش اسمی دارای بالاترین رتبه است. بدین ترتیب یک جدول کامل از کلمات انگلیسی و فارسی و ارتباط آنها توسط فرهنگ لغات مختلف ایجاد شده است.

به دلیل بهبود مجموعه‌های هم‌خانواده در فردوس‌نت، مترادف همه کلمات فارسی موجود از فرهنگ واژگان مترادف و متضاد استخراج شده است. در شکل (۳-۴) مجموعه مترادف‌های کلمه «تعلیم» در سایت www.vajehyab.com نمایش داده شده است.

شکل ۳-۴ : یک نمونه مجموعه مترادف‌های یک کلمه

با توجه به منابع جمع‌آوری شده، مراحل زیر برای ایجاد مجموعه‌های هم‌خانواده در شبکه واژگان فردوس‌نت انجام می‌شود:

- همه کلمات انگلیسی و مجموعه‌های هم‌خانواده شبکه واژگان انگلیسی استخراج و در جداول پایگاه داده فردوس‌نت ذخیره می‌شود.
- اگربرای مجموعه هم‌خانواده شبکه واژگان انگلیسی، مجموعه هم‌خانواده شبکه واژگان فارس‌نت وجود داشته باشد، این مجموعه هم‌خانواده در فردوس‌نت ذخیره می‌شود.
- اگر تعداد کلمات موجود در مجموعه هم‌خانواده فارس‌نت کمتر از تعداد مجموعه هم‌خانواده انگلیسی باشد و یا متناظری برای مجموعه هم‌خانواده انگلیسی در مجموعه

هم‌خانواده فارس‌نت وجود نداشته باشد، کلمات مناسب برای تکمیل این مجموعه هم‌خانواده شناسایی می‌شود.

- ترجمه کلمات مجموعه هم‌خانواده انگلیسی از فرهنگ لغات استخراج می‌شود و برای هر کلمه در مجموعه هم‌خانواده، ترجمه‌ای که بیشترین اشتراک بین ترجمه‌های مختلف کلمات یک مجموعه هم‌خانواده با توجه به رتبه ترجمه، دارا باشد، انتخاب می‌شود.
- اگر مجموعه هم‌خانواده فقط دارای یک کلمه باشد، از توضیحات مجموعه هم‌خانواده استفاده می‌شود. بعد از حذف کلمات ایست از توضیح یک مجموعه هم‌خانواده، آن توضیح به یک سری کلمه شکسته می‌شود. ترجمه‌ی این کلمات در فرهنگ لغات جستجو می‌شود. معانی کلمه مجموعه هم‌خانواده با معانی کلمات توضیح مجموعه هم‌خانواده مقایسه می‌شود. ترجمه‌ای برای کلمه مجموعه هم‌خانواده انتخاب می‌شود که بیشترین اشتراک با معانی کلمات توضیح را داشته باشد.
- کلمات هر مجموعه هم‌خانواده بررسی می‌شوند. اگر کلمات متناظر هر کلمه در مجموعه هم‌خانواده در مجموعه کلمات مترادف آن کلمهوجود داشته باشد، آن مجموعه کلمات مترادف به مجموعه هم‌خانواده اضافه می‌شوند. مثلا اگر کلمه «آموزش» در مجموعه هم‌خانواده «تعلیم» باشد، کلمات «پرورش» و «تربیت» هم به این مجموعه هم‌خانواده اضافه می‌شود.

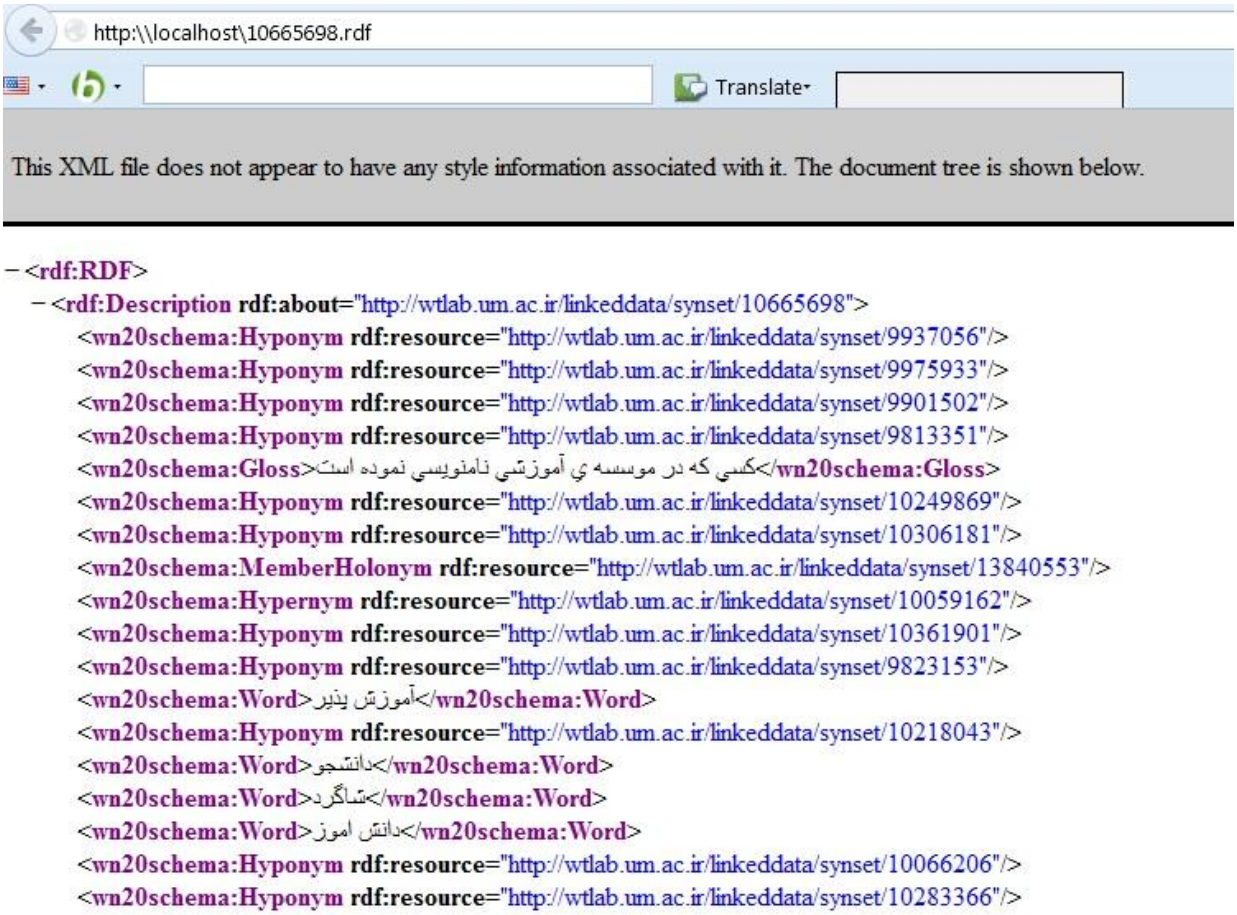
۳-۶-۱- چالش‌ها

در ایجاد خودکار شبکه واژگان فردوس‌نت چالش‌های زیر موجود است:

- بعضی از کلمات و مفاهیم شبکه واژگان انگلیسی در فارسی معادلی ندارند و امکان ترجمه وجود ندارد.
- تعدادی از کلمات شبکه واژگان انگلیسی به عنوان اسم فرد یا اسم محل خاص هستند و قابل ترجمه نیستند.

۳-۷- تولید صفحات RDF از شبکه واژگان

شبکه واژگان فردوس‌نت بر اساس شبکه واژگان انگلیسی می‌باشد. بنابراین روابط بین مجموعه‌های هم‌خانواده در شبکه واژگان فارسی مانند روابط بین مجموعه‌های هم‌خانواده انگلیسی است. این روابط در هستان‌شناسی شبکه واژگان انگلیسی تعریف شده‌اند. در نتیجه برای تولید صفحات RDF از مجموعه‌های هم‌خانواده شبکه واژگان فردوس‌نت، از تعاریف شبکه واژگان انگلیسی استفاده می‌شود. صفحات RDF مجموعه‌های هم‌خانواده شبکه واژگان فردوس‌نت با شماره شناسایی هر مجموعه هم‌خانواده بصورت محلی منتشر شده‌اند. در شکل (۳-۵) صفحه RDF مربوط به یک نمونه مجموعه هم‌خانواده شبکه واژگان فردوس‌نت نمایش داده شده است.



شکل ۳-۵: صفحه RDF برای یک نمونه مجموعه هم‌خانواده

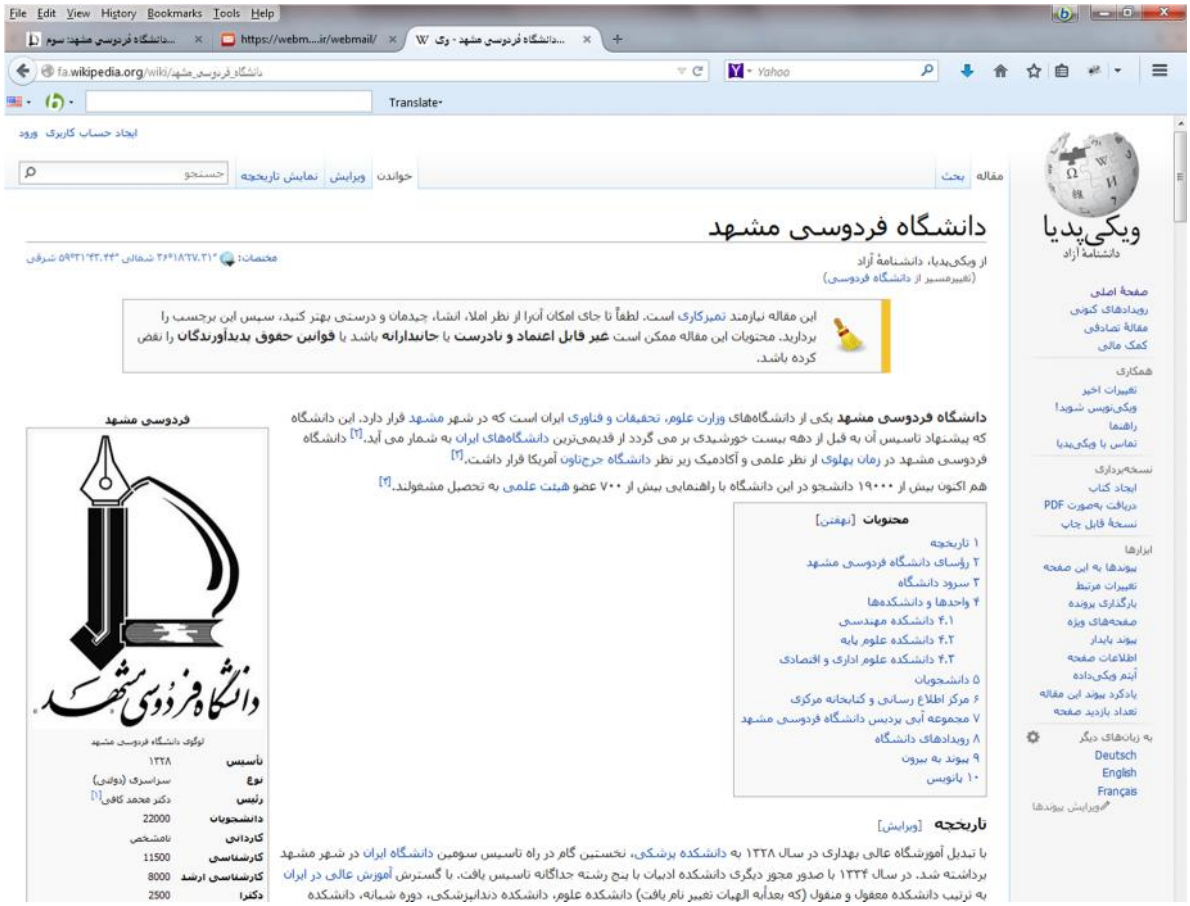
۳-۸- شناسایی لینک برای حقایق

بعد از مرحله استخراج حقایق از سطح جمله به سه‌تایی و آماده‌سازی حقایق برای تبدیل به قالب RDF، لینک مناسب برای قسمت‌های نهاد، مسند و گزاره حقیقت شناسایی می‌شود. در ابتدا جستجو در شبکه واژگان فردوس‌نت انجام می‌پذیرد.

در هر یک از قسمت‌های یک حقیقت یک کلمه اصلی وجود دارد، که کلمات دیگر با استفاده از رابطه‌ی وابستگی به آن کلمه متصل شده‌اند. مثلا کلمه «استاد» در عبارت «استاد دانشگاه مشهد» کلمه اصلی است که برچسب POS آن اسم است و کلمه «کرد» در عبارت «سخنرانی کرد» کلمه اصلی است که برچسب POS آن فعل می‌باشد.

با توجه به برچسب POS کلمه اصلی نهاد، مسند و گزاره حقیقت در بخش‌های اسم، فعل، صفت و قید شبکه واژگان جستجو انجام می‌شود. اگر مجموعه هم‌خانواده‌ای برای هر یک از عبارت‌های نهاد، مسند و گزاره‌ی حقیقت شناسایی شود، آدرس صفحه RDF مربوط به آن مجموعه هم‌خانواده نشان‌دهنده URI آن عبارت حقیقت است.

اگر برای هر یک از عبارت‌های نهاد، مسند و گزاره‌ی حقیقت، مجموعه هم‌خانواده‌ای در شبکه واژگان فردوس‌نت وجود نداشت، جستجو در ویکی‌پدیا انجام می‌شود. اگر صفحه‌ای برای عبارت مورد نظر در ویکی‌پدیا وجود داشت، آدرس آن صفحه بعنوان URI آن عبارت مشخص می‌شود. در شکل (۳-۵) صفحه مربوط به دانشگاه فردوسی نمایش داده شده است.



شکل ۳-۶: صفحه مربوط به دانشگاه فردوسی در ویکی‌پدیا

در قسمت مفروضات سیستم پیشنهادی فرض شده بود که محتوای اسناد مربوط به یک دامنه خاص باشد. در نتیجه جستجو مرتبط با همان دامنه خاص انجام می‌شود.

برای تبدیل حقایق به قالب RDF، برای قسمت‌های نهاد و مسند باید یک URI شناسایی شود، ولی قسمت گزاره می‌تواند مقدار ثابت باشد. در نتیجه اگر لینکی برای قسمت گزاره شناسایی نشود، گزاره بصورت یک رشته بدون لینک ذخیره می‌شود.

۳-۹- خلاصه فصل

در این فصل، روش پیشنهادی پایان‌نامه برای استخراج حقایق از متون فارسی در قالب معنایی ارائه شد. در ابتدا انگیزه‌ی اصلی راه‌کار پیشنهادی برای استفاده از روش استخراج آزاد حقایق در سطح جمله با استفاده از تجزیه وابستگی بیان شد. سپس با معرفی معماری روش پیشنهادی، مراحل آن بطور کامل شرح داده شد.

در مرحله پیش پردازش، متن اسناد به متن به پاراگراف‌ها، جملات، کلمات و عبارات تقسیم می‌شوند. سپس برچسب گذاری POS و تجزیه وابستگی برای جملات انجام می‌شود. در این تجزیه وابستگی، روابط وابستگی بین عناصر جمله مشخص می‌شود. مرحله آخر در پیش‌پردازش کشف مرجع ضمایر و کلمات مانند اختصارات می‌باشد.

در مرحله استخراج حقایق سه‌تایی از جملات، حقایق اصلی و حقایق فرعی جمله استخراج می‌شوند. حقایق اصلی بر اساس افعال جمله و رابطه‌ی آنها با کلمات دیگر جمله شناسایی می‌شوند. در استخراج حقایق فرعی، حقایق مربوط به گروه‌های اسمی مانند مضاف و مضاف‌الیه و موصوف و صفت و همچنین بدل مورد توجه قرار می‌گیرند. در مرحله بعد حقایق برای تبدیل به قالب معنایی RDF آماده می‌شوند. در این مرحله حقایق به حقایق کوچکتر شکسته می‌شوند و همچنین کلمات اضافی حذف می‌شوند. در این مرحله ابعاد زمان و مکان حقایق با توجه به گزاره حقیقت کشف می‌شوند.

برای شناسایی لینک قسمت‌های نهاد، مسند و گزاره حقیقت به یک هستان‌شناسی فارسی نیاز است. در نتیجه در این فصل چگونگی ایجاد خودکار شبکه واژگان فردوس‌نت و انتشار مجموعه‌های هم‌خانواده آن به صورت RDF بصورت محلی توضیح داده شده است. در بخش آخر فصل راه‌کارپیشنهادی، چگونگی شناسایی لینک برای عبارت‌های نهاد، مسند و گزاره حقایق ارائه شده است.

فصل چهارم:

پیاده‌سازی و

ارزیابی

فصل ۴- پیاده‌سازی و ارزیابی

در فصل قبل، جزئیات روش پیشنهادی ارائه گردید. در این فصل، ابتدا روند پیاده‌سازی سیستم پیشنهادی شرح و در ادامه سیستم مورد ارزیابی قرار می‌گیرد.

۴-۱- پیاده‌سازی سیستم پیشنهادی

سیستم پیشنهادی به زبان Java پیاده‌سازی شده است.در این بخش روند پیاده‌سازی سیستم پیشنهادی شرح داده می‌شود.

۴-۱-۱- استخراج حقایق در قالب RDF

در این بخش، بر اساس تجزیه‌ی وابستگی، الگوهای لازم برای استخراج حقایق از جملات شناسایی شد. الگوها با توجه به نوع حقایق به دو دسته تقسیم می‌شوند:

- الگوهای مربوط به استخراج حقایق اصلی جمله بر اساس افعال
- الگوهای مربوط به حقایق فرعی

این الگوها به صورت دستی تهیه شده‌اند. سعی شده است این الگوها در عین سادگی، دارای کارایی بالایی در سیستم پیشنهادی داشته باشند.

در این بخش با استفاده از پیاده‌سازی الگوریتم‌های توضیح داده شده در فصل راه‌کار پیشنهادی، حقایق از جملات (ساده یا مرکب) به صورت سه‌تایی استخراج و برای تبدیل به قالب RDF آماده می‌شوند.

برای شناسایی مکان و زمان حقایق به برچسب‌گذاری نقش معنایی کلمات نیاز است. ولی به علت اینکه SRL با دقت قابل قبولی در فارسی وجود ندارد. در این بخش برای شناسایی مکان و زمان مطابق شکل (۴-۱) از شبکه واژگان استفاده می‌شود.

شکل ۴-۱: شناسایی مکان و زمان با استفاده از شبکه واژگان

با استفاده از کتابخانه Jena در Java، صفحات RDF مجموعه‌های هم‌خانواده شبکه واژگان فردوس‌نت تولید شده است و بصورت محلی منتشر شده است.

در انتها با جستجو در شبکه واژگان و ویکی‌پدیا فارسی لینک برای قسمت‌های نهاد، مسند و گزاره‌ی شناسایی می‌شود. صفحات ویکی‌پدیا بررسی می‌شوند. اگر صفحه‌ای برای عبارت موردنظر موجود باشد، می‌توان محتوای صفحه را خواند، در غیر این صورت صفحه‌ای برای عبارت موردنظر وجود ندارد.

۴-۱-۲- شبکه واژگان فردوس‌نت

برای ایجاد شبکه واژگان فردوس‌نت از شبکه واژگان انگلیسی، شبکه واژگان فارس‌نت، فرهنگ لغات انگلیسی به فارسی و فرهنگ واژگان مترادف و متضاد استفاده شده است. برای پیاده‌سازی شبکه واژگان فردوس‌نت مراحل زیر انجام شده است:

- کلمات انگلیسی همراه با نقش POS آنها و مجموعه‌های هم‌خانواده و روابط بین آنها از شبکه واژگان انگلیسی استخراج شده‌اند. پایگاه داده شبکه واژگان فردوس‌نت بر اساس شبکه واژگان انگلیسی ایجاد و تنظیم شده است.
- کلمات و مجموعه‌های هم‌خانواده و روابط بین آنها از فایل‌های XML فارس‌نت استخراج و در پایگاه داده ذخیره شده‌اند.
- ترجمه‌ی کلمات شبکه واژگان انگلیسی با استفاده از فرهنگ لغت Translate Google جمع‌آوری شده و در پایگاه داده ذخیره شده است. همان گونه که در فصل راه‌کار پیشنهادی مطرح شد، ترجمه‌ی هر کلمه با توجه به نقش POS و رتبه آن در لیست کلمات ترجمه ذخیره می‌شود.
- لغت‌نامه‌های Babylon، Tarjomeh.com، Farsi123.com و Dic.abadis.net بررسی شده‌اند و تغییرات زیر بر روی آنها اعمال شده است.
    - جداسازی معنایی مختلف یک کلمه
    - جدا کردن توضیحات اضافی در یک معنی
    - حذف معانی دارای کلمه انگلیسی (اگر کلمه انگلیسی در توضیح اضافی باشد، مشکلی ندارد)
    - دادن رتبه به معانی

- یک جدول کامل از کلمات انگلیسی و فارسی و ارتباط آنها توسط لغت‌نامه‌ها ایجاد شده است.
- مترادف‌ها و متضادها همه کلمات فارسی بدست آمده در مراحل قبل با استفاده از فرهنگ واژگان مترادف و متضاد سایت www.vajehyab.com استخراج شده است.
- در انتها الگوریتم پیشنهادی برای ایجاد شبکه واژگان فردوس‌نت بر روی پایگاه داده‌های جمع‌آوری شده مراحل قبل اعمال شد.

۴-۲- ارزیابی

در این بخش بعد از معرفی مجموعه داده و معیارهای ارزیابی، سیستم پیشنهادی در بخش‌های تولید حقایق به صورت سه‌تایی، ایجاد شبکه واژگان و شناسایی لینک برای قسمت‌های نهاد، مسند و گزاره‌ی یک حقیقت مورد مقایسه و ارزیابی قرار می‌گیرد و نتایج هر بخش ارائه می‌شود.

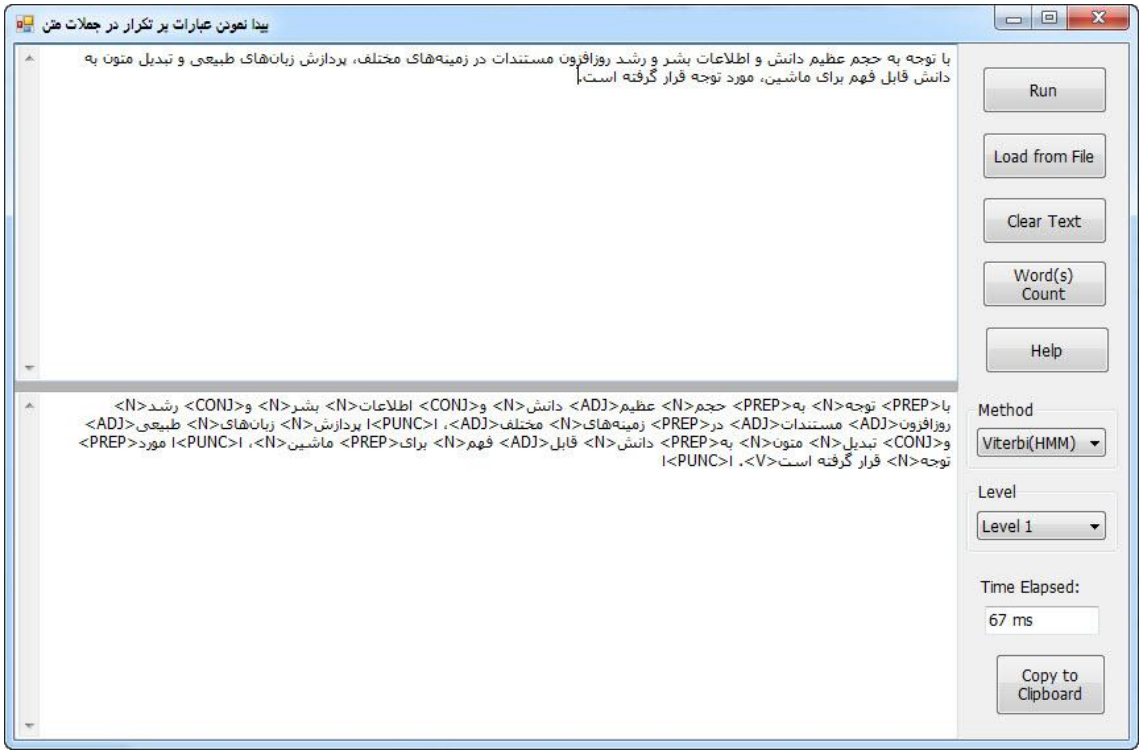
۴-۲-۱- مجموعه داده

همانطور که در فصل راه‌کار پیشنهادی مطرح شد، ورودی سیستم پیشنهادی جملات برچسب‌خورده است. مجموعه داده دادگان شامل جملاتی می‌باشد که بصورت دستی برچسب‌گذاری شده‌اند. این مجموعه داده بر اساس قالب معیار همایش زبان‌شناسی رایانه‌ای و پردازش زبان طبیعی است. توضیحات در مورد اختصارات موجود در این مجموعه داده در جداول پیوست نشان داده شده است. برای نمونه جمله «فیلم‌بردار و خواهرش در حالی که جیغ می‌کشند، شروع به دویدن می‌کنند.» در مجموعه داده دادگان برچسب‌گذاری شده است:

1	فیلم‌بردار	فیلم‌بردار	N	ANM	attachment=ISO|number=SING|senID=35126	13	SBJ	_	_
2	و	و	CONJ	CONJ	attachment=ISO|senID=35126	1	NCONJ	_	_
3	خواهرش	خواهر	N	ANM	attachment=ISO|number=SING|senID=35126	2			

	POSDEP	_	_						
4	در	در	PREP	PREP	attachment=ISO|senID=35126	13	ADV	_	_
5	حالی	حال	N	IANM	attachment=ISO|number=SING|senID=35126	4	POSDEP		
	_	_							
6	که	که	SUBR	SUBR	attachment=ISO|senID=35126	5	NCL	_	_
7	جیغ	جیغ	N	IANM	attachment=ISO|number=SING|senID=35126	8	NVE	_	
	_								
8	می‌کشند	کشید#کش	V	ACT					
	person=3|attachment=ISO|number=PLUR|tma=H|senID=351266						PRD	_	_
9	،	،	PUNC	PUNC	attachment=ISO|senID=35126	8	PUNC	_	_
10	شروع	شروع	N	IANM	attachment=ISO|number=SING|senID=35126	13	NVE	_	
	_								
11	به	به	PREP	PREP	attachment=ISO|senID=35126	10	NPP	_	_
12	دویدن	دویدن	N	IANM	attachment=ISO|number=SING|senID=35126	11	POSDEP		
	_	_							
13	می‌کنند	کرد#کن	V	ACT					
	person=3|attachment=ISO|number=PLUR|tma=H|senID=351260						ROOT	_	_
14	.	.	PUNC	PUNC	attachment=ISO|senID=35126	13	PUNC	_	_

در این سیستم از پیکره دو زبانه فارسی- انگلیسی آزمایشگاه فناوری وب هم بعنوان ورودی سیستم پیشنهادی استفاده شده است. برچسب‌گذاری POS جملات فارسی با استفاده از ابزار پردازش زبان طبیعی آزمایشگاه فناوری وب دانشگاه فردوسی مانند شکل (۴-۲) انجام می‌شود. برچسب‌گذاری تجزیه‌ی وابستگی جملات به صورت دستی با همکاری اساتید زبان‌شناسی بر اساس پیکره دادگان انجام شده است.



شکل ۴-۲: ابزار پردازش زبان طبیعی آزمایشگاه فناوری وب دانشگاه فردوسی

۴-۲-۲- معیارهای ارزیابی سیستم پیشنهادی

کارایی سیستم‌های استخراج اطلاعات با دو معیار دقت و فراخوانی سنجیده می‌شوند. دقت، نسبت تعداد حقایق صحیح استخراج شده به تعداد کل حقایق شناسایی شده است. فراخوانی، نسبت تعداد حقایق صحیح استخراج شده به تعداد کل حقایق موجود می‌باشد. همچنین برای ترکیب این دو، معیار F_Measure هم به صورت فرمول (۴-۱) تعریف می‌شود:

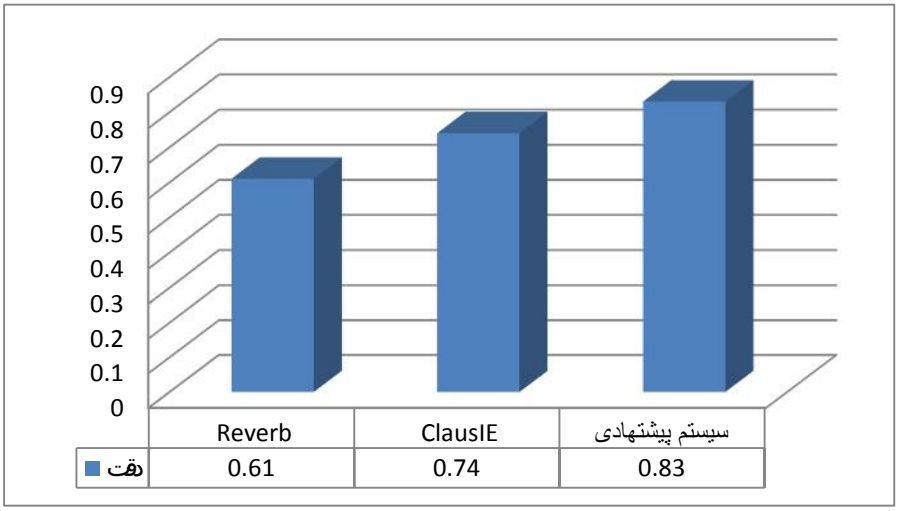
$$F\_Measure = 2\times\frac{Precision\times Recall}{Precision + Recall} \qquad (۱-۴)$$

معمولا در سیستم‌های استخراج اطلاعات میزان فراخوانی کمتر از مقدار دقت می‌باشد. با توجه به تعاریف آنها مشخص است که مقادیر آنها با یکدیگر مرتبط می‌باشد. اغلب روش‌هایی که برای بالا بردن یک معیار استفاده می‌شود، باعث کاهش معیار دیگر می‌شود ولی در نتیجه‌ی معیار F_Measure

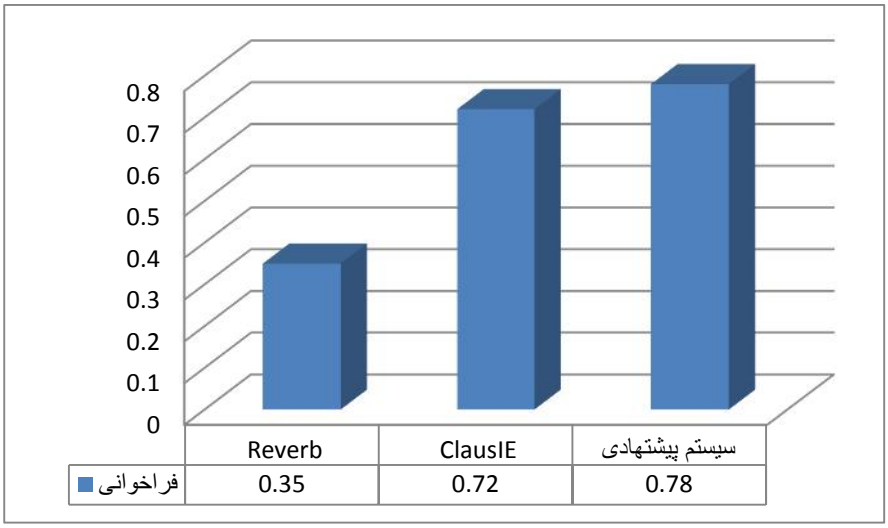
تغییر چندانی اعمال نمی‌شود. در این تحقیق سعی شده است با استفاده از تجزیه وابستگی، الگوها و الگوریتم‌های بخش‌های مختلف راه‌کار پیشنهادی هر دو معیار بهبود یابند.

۴-۲-۳- مقایسه و ارزیابی سیستم پیشنهادی در بخش تولید حقایق به صورت سه‌تایی

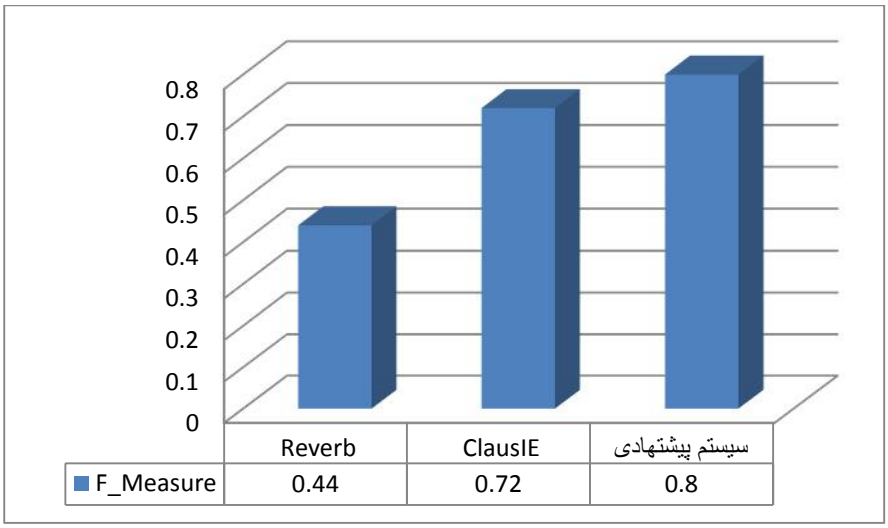
در بین سیستم‌های ذکر شده در بخش پیشینه،سیستم پیشنهادی در راستای سیستم‌های ClausIE [12] و Reverb [11] است. در نتیجه این سیستم‌ها جهت مقایسه و ارزیابی انتخاب شده‌اند. سیستم مرصاد که توسط دانشگاه شهید بهشتی برای زبان فارسی ارائه شده است محدود به الگوهای ثابت و افعال خاص می‌باشد، در نتیجه برای ارزیابی سیستم پیشنهادی از آن استفاده نشده است. در شکل‌های (۴-۳)، (۴-۴) و (۴-۵) نتایج ارزیابی و مقایسه سیستم پیشنهادی نمایش داده شده است. در این بخش از جملات پیکره دو زبانه فارسی- انگلیسی آزمایشگاه فناوری وب استفاده شده است.



شکل ۴-۳: ارزیابی و مقایسه دقت سیستم پیشنهادی



شکل ۴ -۴: ارزیابی و مقایسه فراخوانی سیستم پیشنهادی



شکل ۴ -۵: ارزیابی و مقایسه F_Measure سیستم پیشنهادی

همانطور که در بخش کارهای پیشین بیان شد، سیستم ClausIE دقت و فراخوانی بالاتری نسبت به سایر سیستم‌های استخراج آزاد دیگر دارد. ولی یکی از چالش‌های این سیستم، کند بودن آن است. کارایی سیستم به تجزیه‌گر استنفورد وابسته است. بعضی از حقایقی که در این سیستم استخراج می‌شوند، با یکدیگر هم‌پوشانی دارند. از بعضی جملات، حقایق اضافی و اشتباه بدست می‌آید، مثلا صفت ملکی به تنهایی در قسمت نهاد قرار می‌گیرد.

اساس کار سیستم Reverb با افعال است، در نتیجه فراخوانی آن کاهش می‌یابد. در بعضی حقایق، نهاد و گزاره صحیح تشخیص داده نمی‌شود، که باعث کاهش دقت سیستم می‌شود. ورودی سیستم پیشنهادی جملات برچسب خورده است. دقت سیستم پیشنهادی در تولید حقایق سه‌تایی تحت تاثیر چالش‌هایی مانند شناسایی نهاد و گزاره حقایق در جملات پیچیده می‌باشد. البته در سیستم پیشنهادی حقایق برای تبدیل به مدل داده‌ای RDF آماده‌سازی می‌شوند، در نتیجه حقایق تا حد امکان شکسته و کوچک می‌شوند.در جدول (۴-۱) سیستم پیشنهادی از نظر کیفی با سایر سیستم‌های استخراج آزاد و همچنین سیستم مرصاد مقایسه شده است.

جدول ۴-۱: مقایسه کیفی سیستم پیشنهادی با سایر سیستم‌های استخراج آزاد

سیستم	استخراج روابط بامعنی ممکن	لینک به توضیحات معنایی	عدم وابستگی به گرامر زبان	نبود استخراج‌های غیرحقیقی	عدم استخراج روابط بی‌معنی و یا خالی از اطلاعات مفید	عدم نیاز به داده آموزشی
WOE [7]	✓		✓			
StatSnowball [3]	✓		✓			
ReVerb [4]	✓				✓	✓
R2A2 [7]	✓				✓	
OLLIE [8]	✓			✓	✓	
ClauseIE [6]	✓				✓	✓
سیستم مرصاد [14]					✓	✓
سیستم پیشنهادی	✓	✓			✓	✓

۴-۲-۴- مقایسه و ارزیابی شبکه واژگان فردوس‌نت

در این بخش شبکه واژگان فردوس‌نت با شبکه واژگان انگلیسی و شبکه واژگان فارس‌نت مقایسه شده است. در ابتدا مقایسه‌ی آماری بر روی سه شبکه واژگان انگلیسی، فارس‌نت و فردوس‌نت انجام می‌شود. در ادامه دقت و فراخوانی شبکه واژگان فارس‌نت با شبکه واژگان فردوس‌نت مقایسه می‌شود.

دقت، نسبت تعداد کلمات مناسب هر مجموعه هم‌خانواده به تعداد کل کلمات شناسایی شده است. فراخوانی، نسبت تعداد کلمات مناسب بدست آمده به تعداد کل کلمات موجود در یک مجموعه هم‌خانواده از نظر فرد خبره می‌باشد.

در جدول (۴-۲) تعداد مجموعه‌های هم‌خانواده در دسته‌های اسم، فعل، صفت و قید در شبکه واژگان فردوس‌نت نمایش داده شده است و با شبکه واژگان انگلیسی و شبکه واژگان فارس‌نت مقایسه شده است.

جدول ۴-۲: مقایسه شبکه واژگان فردوس‌نت از نظر تعداد مجموعه‌های هم‌خانواده

مجموعه هم‌خانواده	شبکه واژگان انگلیسی	شبکه واژگان فارس‌نت	شبکه واژگان فردوس‌نت
اسم	۸۲،۱۱۵	۲۲،۲۱۹	۶۹،۹۷۱
فعل	۱۳،۷۶۷	۶،۳۶۶	۱۳،۵۱۱
صفت	۱۸،۱۵۶	۵،۲۵۰	۱۶،۴۸۰
قید	۳،۶۲۱	۲۲۴	۳،۱۵۱
جمع	۱۱۷،۶۵۹	۳۴،۰۵۹	۱۰۳،۱۱۳

شبکه‌ی واژگان فردوس‌نت بر اساس شبکه واژگان انگلیسی است؛ اگر برای کلمات هر مجموعه هم‌خانواده در شبکه واژگان انگلیسی با توجه به راه‌کار پیشنهادی، معادلی در فارسی شناسایی شود، مجموعه هم‌خانواده مورد نظر در شبکه واژگان فردوس‌نت اضافه می‌شود.

در شبکه واژگان فردوس‌نت برای شناسایی کلمات هر مجموعه هم‌خانواده، علاوه بر شبکه واژگان فارس‌نت از فرهنگ لغات انگلیسی به فارسی و فرهنگ واژگان مترادف و متضاد استفاده شده است.جدول (۴-۳) مربوط به تعداد کلمات موجود در دسته‌های اسم، فعل، صفت و قید در شبکه واژگان انگلیسی، شبکه واژگان فردوس‌نت و شبکه واژگان فارس‌نت است.

جدول ۴-۳: مقایسه شبکه واژگان فردوس‌نت از نظر تعداد عبارات در مجموعه‌های هم‌خانواده

مجموعه هم‌خانواده	شبکه واژگان انگلیسی	شبکه واژگان فارس‌نت	شبکه واژگان فردوس‌نت
اسم	۱۱۷،۷۹۸	۱۷،۰۲۵	۱۴۲،۷۳۲
فعل	۱۱،۵۲۹	۵،۶۶۶	۲۴،۵۷۰
صفت	۲۱،۴۷۹	۶،۵۴۳	۳۱،۸۰۲
قید	۴،۴۸۱	۲،۰۱۵	۶،۲۷۵
جمع	۱۴۷،۳۰۶	۲۹،۸۹۲	۱۸۴،۸۰۴

همانگونه در جدول (۴-۳) مشاهده می‌شود، تعداد عبارت‌های موجود در شبکه واژگان فارسی بیشتر از شبکه واژگان انگلیسی می‌باشد، زیرا برای یک کلمه انگلیسی حداکثر ترجمه‌های مختلفی که با کلمات دیگر مجموعه هم‌خانواده مشترک هستند، انتخاب شده است. در نتیجه ممکن است برای یک کلمه انگلیسی چندین معادل فارسی در مجموعه هم‌خانواده شناسایی شود. علاوه بر این کلماتی هم از مجموعه واژگان مترادف اضافه می‌شود.

در جدول (۴-۴) تعداد انواع روابط بین مجموعه‌های هم‌خانواده در دسته‌های مختلف اسم، فعل، صفت و قید در شبکه واژگان انگلیسی، شبکه واژگان فارس‌نت و شبکه واژگان فردوس‌نت آورده شده است.

جدول ۴-۴: مقایسه شبکه واژگان فردوس‌نت از نظر تعداد انواع روابط بین مجموعه‌های هم‌خانواده

مجموعه هم‌خانواده	شبکه واژگان انگلیسی	شبکه واژگان فارس‌نت	شبکه واژگان فردوس‌نت
اسم	۱۷	۲۴	۱۷
فعل	۹	۱۲	۹
صفت	۶	۱۳	۶
قید	۳	۸	۳
جمع	۲۲	۳۲	۲۲

روابط موجود در شبکه واژگان فردوس‌نت مطابق روابط موجود در شبکه واژگان انگلیسی است. در شبکه واژگان فارس‌نت، روابط مفهومی بین مجموعه‌های هم‌خانواده بر اساس تولید نیمه‌خودکار شبکه مانند روش مبتنی بر الگو و روش آماری شناسایی می‌شوند. روابط مفهومی شبکه واژگان فارس‌نت شامل موارد زیر می‌باشد:

Hypernym	agent عامل
Hyponym	patient ذیرا
Holonym(part of)	is-Caused-by
Meronym(part of)	is-Instrument-of
Holonym(member of)	has-attribute
related-to	Synonym
Antonym	instrument ابزار
Holonym(portion of)	See also
Meronym(portion of)	Has Holo made of

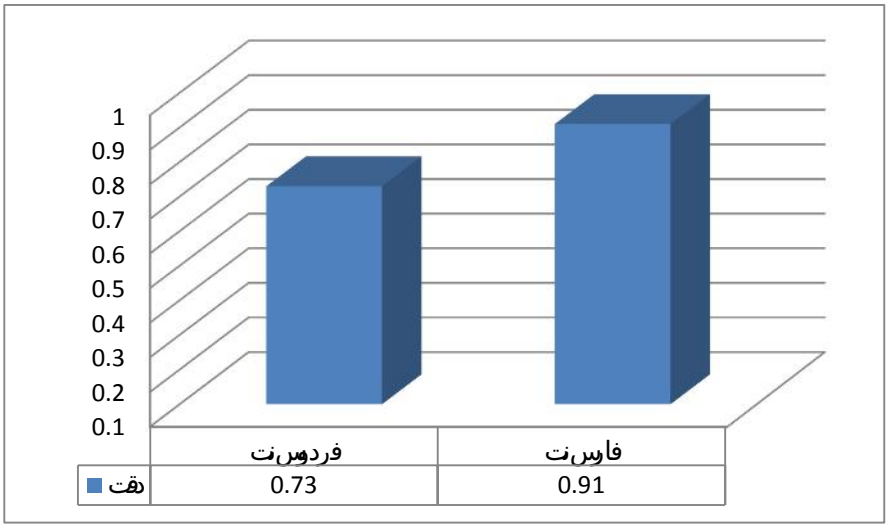
Causes	is-Entailed-by
Derivational Form	Is-Domain-of
co-occurrence	Domain
Meronym(member of)	is-Agent-of
Has Mero made of	is-attribute-of
coordinate term	is-Patient-of
صفت برجسته است برای	دارای صفت برجسته

در جدول(۴-۵) دقت شبکه واژگان فردوس‌نت با شبکه واژگان فارس‌نت در دسته‌های اسم، فعل، صفت و قید مقایسه شده است. برای ارزیابی و مقایسه ۳۰ مجموعه هم‌خانواده برای مجموعه‌خانواده‌های اسم و فعل و ۲۰ مجموعه هم‌خانواده از مجموعه‌خانواده‌های صفت و قید بصورت تصادفی از شبکه واژگان فردوس‌نت و شبکه واژگان فارس‌نت انتخاب شده‌اند. در ارزیابی دقت مجموعه‌های هم‌خانواده از کلمات و توضیحات مجموعه‌های هم‌خانواده شبکه واژگان انگلیسی و فرهنگ واژگان مترادف و متضاد فارسی استفاده می‌شود.

جدول ۴-۵: مقایسه دقت مجموعه‌های هم‌خانواده شبکه واژگان فردوس‌نت با فارس‌نت

مجموعه هم‌خانواده	شبکه واژگان فارس‌نت	شبکه واژگان فردوس‌نت
اسم	۰,۸۸	۰,۶۳
فعل	۰,۹۷	۰,۷۲
صفت	۰,۹۵	۰,۷۹
قید	۰,۹۴	۰,۸۰

در ادامه بصورت تصادفی ۳۰ مجموعه هم‌خانواده از شبکه واژگان فارس‌نت و شبکه واژگان فردوس‌نت انتخاب شده‌اند و دقت مجموعه هم‌خانواده‌های این دو شبکه واژگان در شکل (۴-۶) مقایسه شده‌اند.



شکل ۴-۶: مقایسه دقت مجموعه‌های هم‌خانواده‌ی شبکه واژگان فردوس‌نت با فارس‌نت

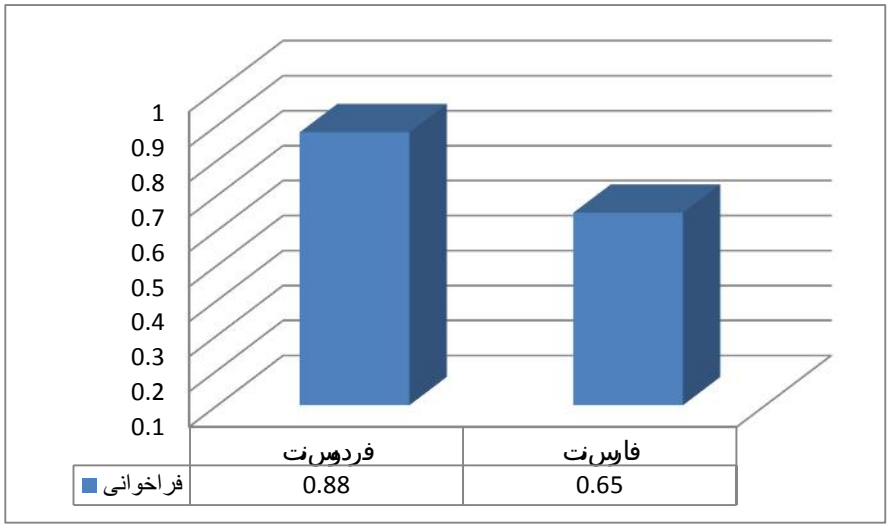
در جدول (۴-۶) فراخوانی شبکه واژگان فردوس‌نت با شبکه واژگان فارس‌نت مقایسه شده است.

جدول ۴-۶: مقایسه فراخوانی مجموعه‌های هم‌خانواده شبکه واژگان فردوس‌نت با فارس‌نت

مجموعه هم‌خانواده	شبکه واژگان فارس‌نت	شبکه واژگان فردوس‌نت
اسم	۰,۵۸	۰,۸۳
فعل	۰,۷۰	۰,۸۹
صفت	۰,۶۸	۰,۹۰
قید	۰,۷۱	۰,۹۵

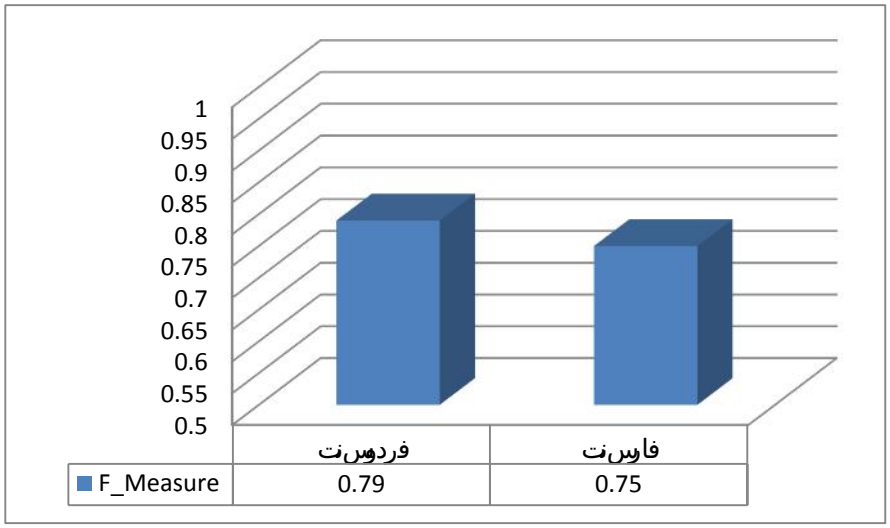
علاوه بر این که تعداد مجموعه‌های هم‌خانواده شبکه واژگان فردوس‌نت خیلی بیشتر از شبکه واژگان فارس‌نت می‌باشد، همچنین مجموعه‌های هم‌خانواده شبکه واژگان فردوس‌نت دارای فراخوانی بالاتری نسبت به شبکه واژگان فارس‌نت هستند.

برای محاسبه فراخوانی کلی مجموعه‌های هم‌خانواده‌ها، بصورت تصادفی ۳۰ مجموعه هم‌خانواده از شبکه واژگان فارس‌نت و شبکه واژگان فردوس‌نت انتخاب شده‌اند. فراخوانی مجموعه هم‌خانواده‌های این دو شبکه واژگان در شکل (۴-۷) مقایسه شده‌اند.



شکل ۴ -۷: مقایسه فراخوانی مجموعه‌های هم‌خانواده‌ی شبکه واژگان فردوس‌نت با فارس‌نت

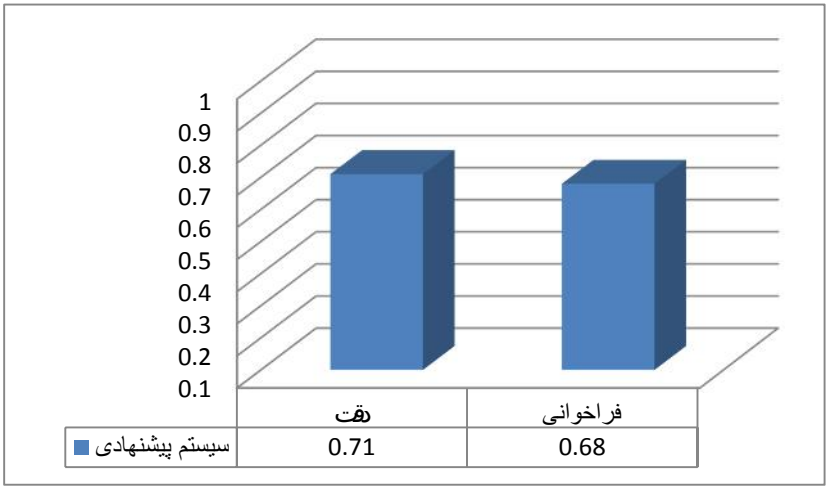
درشکل (۴-۸) معیارF_Measureشبکه واژگان فردوس‌نت و شبکه واژگان فارس‌نت محاسبه و مقایسه شده است.



شکل ۴ -۸: مقایسه معیار F_Measure مجموعه‌های هم‌خانواده‌ی شبکه واژگان فردوس‌نت با فارس‌نت

۴-۲-۵- ارزیابی حقایق در قالب RDF

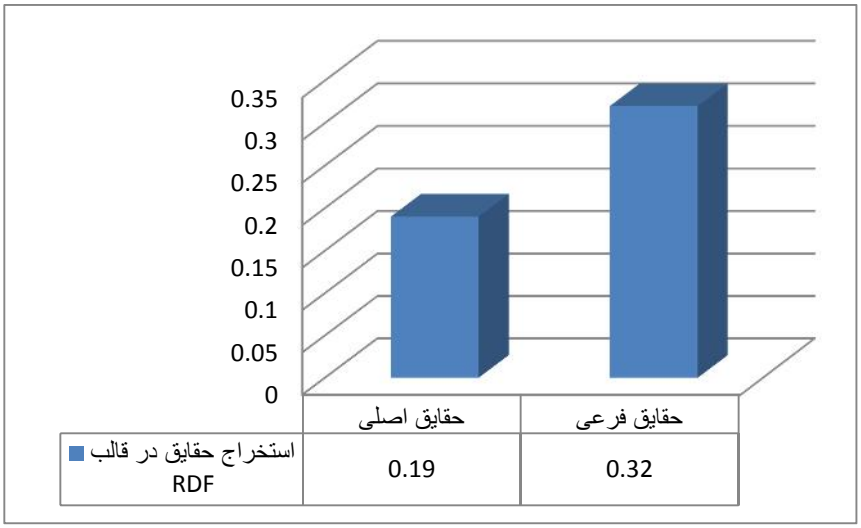
در این بخش، از جملات مجموعه داده دادگان استفاده شده است. در شکل (۴-۷) دقت و فراخوانی استخراج حقایق با استفاده از سیستم پیشنهادی بر روی مجموعه داده دادگان نشان داده شده است.



شکل ۴-۹: دقت و فراخوانی سیستم پیشنهادی بر روی جملات مجموعه داده دادگان

دقت و فراخوانی سیستم پیشنهادی به پیچیدگی جملات پیکره مورد استفاده وابسته است. در نتیجه اگر پیکره دارای جملات پیچیده‌تری باشد، سیستم پیشنهادی دارای دقت و فراخوانی پایین‌تر است.

در این بخش ارزیابی با توجه به بخش شناسایی لینک برای حقایق که در فصل راه‌کار پیشنهادی توضیح داده شده است، انجام می‌شود. در نتیجه اگر برای قسمت‌های نهاد و مسند حقیقت یک URI شناسایی شود، این حقیقت می‌تواند در قالب معنایی RDF نگاشت شود. نتایج ارزیابی سیستم پیشنهادی در استخراج حقایق در قالب معنایی RDF در شکل (۴-۸) نمایش داده شده است.



شکل ۴-۱۰: ارزیابی سیستم پیشنهادی در نگاشت حقایق در قالب معنایی RDF

همانطور که در شکل (۴-۸) مشاهده می‌شود، حقایق فرعی بیشتر از حقایق اصلی در قالب معنایی RDF نگاشت می‌شوند. حقایق فرعی نسبت به حقایق اصلی کوچکتر هستند، همچنین نهاد این حقایق دارای عبارات ساده‌تری می‌باشند. از طرفی برای مسند حقایق فرعی بر اساس راه‌کار پیشنهادی از یک مجموعه افعال ثابت استفاده شده است.

برای مثال برای نهاد حقایق اصلی زیر لینکی شناسایی نشده است.

- («رئیس بیمه مرکزی جمهوری اسلامی ایران »، «خاطرنشان کرد »، «بر اساس آخرین آمار استخراج شده ضریب نفوذ بیمه ۱ درصد است»)
- («ضریب نفوذ بیمه »، «است»، «۱ درصد »)
- («ضریب نفوذ بیمه »، «است»، «بر اساس آخرین آمار استخراج شده »)

حقایق اصلی زیر در قالب معنایی RDF نگاشت شده‌اند.

- («سال نو»، «آمدن»،)

(http://wtlab.um.ac.ir/linkeddata/synset/15182189.rdf,
http://wtlab.um.ac.ir/linkeddata/synset/1849221.rdf, BlankNode)

- («سال نو»، «نشستن»، « کنار سفرهٔ هفت‌سین»)

(http://wtlab.um.ac.ir/linkeddata/synset/15182189.rdf,
http://wtlab.um.ac.ir/linkeddata/synset/1979901.rdf, ‘کنار سفرهٔ هفت‌سین’)

۴-۳- خلاصه فصل

در این فصل، در ابتدا نحوه پیاده‌سازی سیستم پیشنهادی شرح داده شده است. پیاده‌سازی سیستم پیشنهادی در دو بخش ایجاد خودکار شبکه واژگان فردوس‌نت و استخراج حقایق در قالب RDF بیان شده است. همچنین مجموعه داده مورد استفاده در سیستم پیشنهادی معرفی شده است. سپس در بخش ارزیابی بعد از بیان معیارهای ارزیابی سیستم پیشنهادی در سه بخش تولید حقایق سه‌تایی، ایجاد شبکه واژگان فردوس‌نت و تولید حقایق در قالب RDF با سیستم‌های دیگر مقایسه و ارزیابی شده است. نتایج حاصل از ارزیابی نشان می‌دهد که روش پیشنهادی در استخراج حقایق و ایجاد شبکه واژگان فردوس‌نت موفق بوده و نتایج مطلوبی را بدست آورده است.

فصل پنجم:

نتیجه‌گیری و کارهای آینده

فصل ۵- نتیجه‌گیری و کارهای آینده

۵-۱- نتیجه‌گیری

در جهت بهره‌برداری و استفاده بهتر از محتوای اسناد و صفحات وب، اطلاعات آنها به دانش قابل فهم برای ماشین تبدیل می‌شود. از این طریق می‌توان با استفاده از زبان‌های پرس‌وجوگر بر روی پایگاه دانش به اطلاعات مفید و جامعی دست یافت. سیستم‌های استخراج اطلاعات با پردازش متون اسناد، حقایق را در قالب‌های ساخت‌یافته استخراج می‌کنند. با توجه به انواع مختلف سیستم‌های استخراج اطلاعات و چالش‌های آنها، راه‌کار پیشنهادی سعی در استخراج حقایق برای نگاشت به قالب معنایی RDF از متون فارسی را دارد. در این راستا از تجزیه وابستگی و ارتباط وابستگی بین کلمات جمله استفاده می‌کند.

سیستم پیشنهادی حقایق را در دسته‌های متفاوت با توجه به اهمیت آنها کشف می‌کند. حقایق اصلی و حقایق فرعی جمله با توجه به تجزیه وابستگی شناسایی می‌شوند. در راه‌کار پیشنهادی، حقایق برای تبدیل به سه‌گانه‌های قالب معنایی RDF آماده می‌شوند. جنبه دیگر راه‌کار پیشنهادی کشف ابعاد زمان و مکان یک حقیقت می‌باشد.

در راستای شناسایی URI برای قسمت‌های نهاد، مسند و گزاره حقایق در فارسی به یک هستان‌شناسی کامل نیاز است. در نتیجه در راه‌کارپیشنهادی، شبکه واژگان فردوس‌نت به صورت خودکار ایجاد شده است. شبکه واژگان فردوس‌نت بر پایه‌ی شبکه واژگان انگلیسی است. برای ایجاد شبکه واژگان فردوس‌نت از مجموعه‌های هم‌خانواده شبکه واژگان فارس‌نت استفاده شده است. برای ترجمه کلمات هر مجموعه هم‌خانواده، فرهنگ‌لغات مختلف جمع‌آوری و استفاده شده‌اند.

مجموعه‌های هم‌خانواده شبکه واژگان فردوس‌نت به صورت فایل‌های RDF بصورت محلی منتشر شده‌اند. در بخش شناسایی URI از صفحات RDF شبکه واژگان و ویکی‌پدیا استفاده شده است.

همانگونه که در فصل ارزیابی ارائه شد، سیستم پیشنهادی در استخراج حقایق موفق بوده و باعث بهبود دقت و فراخوانی نسبت به سیستم‌های موجود شده است. همچنین شبکه واژگان فردوس‌نت با استفاده از راه‌کار پیشنهادی برای ایجاد خودکار شبکه واژگان، با دقت خوبی تولید شده است. سیستم پیشنهادی حقایق را برای نگاشت به مدل داده‌ای RDF استخراج می‌کند. در نتیجه می‌توانبا استفاده از لینک به داده‌های پیوندی و سایر هستان‌شناسی‌ها و توضیحات معنایی اطلاعات بیشتری در مورد آرگومان‌ها و رابطه حقایق بدست آورد.

۵-۲- کارهای آتی

تحلیل معنایی بیشتر حقایق: همانطور که در چالش‌های مطرح شده در فصل راه‌کار پیشنهادی بیان شد، اگر تحلیل معنایی بیشتری بر روی حقایق انجام شود، در نتیجه حقایق با کیفیت بهتری برای تبدیل به قالب معنایی RDF آماده می‌شوند.

ابهام‌زدایی موجودیت‌ها و روابط: با توجه به محتوای اسناد، برای شناسایی URI مناسب برای موجودیت‌ها و روابط ابهام‌زدایی انجام شود.

مقیاس‌پذیری: با توجه به این که استخراج اطلاعات و برنامه‌های کاربردی متن کاوی با مقدار بسیار زیادی از اطلاعات متنی با ارزش روبرو هستند، و از طرفی استخراج موجودیت‌ها و روابط از اسناد از نظر محاسباتی یک کار سنگین و گران می‌باشد. اگر مقیاس‌پذیری سیستم استخراج اطلاعات ارتقا یابد، می‌توان اطلاعات بیشتری از اسناد در زمان کمتری بدست آورد.

ارزیابی خودکار سیستم استخراج اطلاعات: همانگونه که در فصل مرور ادبیات مطرح شد، ارزیابی حقایق استخراج شده را می‌توان از دو جنبه بررسی نمود؛ اول اینکه صحت و درستی حقایق استخراج شده مورد ارزیابی قرار گیرد و جنبه دوم مربوط به ارزشمندی این حقایق می‌باشد. با توجه به استخراج خودکار حقایق، معیار سنجش صحت و ارزشمندی یکی از چالش‌ها در ارزیابی اطلاعات در استخراج آزاد می‌باشد. در حال حاضر ارزیابی سیستم‌های استخراج اطلاعات با استفاده از افراد خبره و یا میزان کارایی کاربرد آنها در سیستم‌های دیگر مورد بررسی قرار می‌گیرد.

تولید تجزیه‌گر وابستگی با دقت بالا: اگر تجزیه‌گر وابستگی با دقت بالایی تهیه شود، می‌توان از آن در سیستم پیشنهادی استفاده نمود.

منابع

منابع

[1] G. Weikum, M. Theobald, “From information to knowledge: harvesting entities and relationships from web sources”, In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 65-76, ACM, 2010.

[2] O. Etzioni, M. Banko, S. Soderland, D. S. Weld, “Open information extraction from the web”,*Communications of the ACM*, *51*(12), 68-74, 2008.

[3] J. Zhu, Z. Nie, X. Liu, B. Zhang, J. R. Wen, “StatSnowball: a statistical approach to extracting entity relationships”, In *Proceedings of the 18th international conference on World wide web*, pp. 101-110, ACM, 2009.

[4] A. Fader, S. Soderland, O. Etzioni, “Identifying relations for open information extraction”, In *Proceedings of the Conference on Empirical Methods in Natural*

Language Processing, pp. 1535-1545, Association for Computational Linguistics, 2011.

[5] B. Hannah, E. Haussmann, "More Informative Open Information Extraction via Simple Inference" Advances in Information Retrieval, Springer International Publishing, 585-590, 2014.

[6] L. Del Corro, R. Gemulla, “ClausIE: clause-based open information extraction”, In *Proceedings of the 22nd international conference on World Wide Web*, pp. 355-366, International World Wide Web Conferences Steering Committee, 2013.

[7] O. Etzioni, A. Fader, J. Christensen, S. Soderland, M. Mausam, “Open information extraction: The second generation”, In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume One*, pp. 3-10, AAAI Press, 2011.

[8] M. Schmitz, R. Bart, S. Soderland, O. Etzioni, “Open language learning for information extraction”. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 523-534, Association for Computational Linguistics, 2012.

[9] F. M. Suchanek, G. Kasneci, G. Weikum, “Yago: a core of semantic knowledge”, In Proceedings of the 16th international conference on World Wide Web, pp. 697-706, ACM, 2007.

[10] J. Biega, E. Kuzey, F. M. Suchanek, “Inside YOGO2s: a transparent information extraction architecture”, In *Proceedings of the 22nd international conference on World Wide Web companion*, pp. 325-328, International World Wide Web Conferences Steering Committee, 2013.

[11] J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. De Melo, G. Weikum, “Yago2: exploring and querying world knowledge in time, space, context, and many languages”, In *Proceedings of the 20th international conference companion on World wide web*, pp. 229-232, ACM, 2011.

[12] M. J. Cafarella, D. Downey, S. Soderland, O. Etzioni, “KnowItNow: Fast, scalable information extraction from the web”, In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 563-570, Association for Computational Linguistics, 2005.

[13] C. H. Chang, M. Kayed, R. Girgis, K. F. Shaalan, “A survey of web information extraction systems”,*Knowledge and Data Engineering, IEEE Transactions on*, *18*(10), 2006.

[۱۴] ن. رحیمی‌پور، م .شمس‌فرد، ز.انصاری، "سیستم استخراج اطلاعات: مرصاد"، پانزدهمین کنفرانس مهندسی برق ایران، تهران، مرکز تحقیقات مخابرات ایران،۱۳۸۶.

[15] T. Hishiki, N. Collier, C. Nobata, T. Okazaki-Ohta, N. Ogata, T. Sekimizu, “Developing NLP Tools for Genome Informtics: An Information Extraction Perspective”,*GENOME INFORMATICS SERIES*, 81-90, 1998.

[16] B. Libbus, T. C. Rindflesch, “NLP-based information extraction for managing the molecular biology literature”, In *Proceedings of the AMIA Symposium*, p. 445, American Medical Informatics Association, 2002.

[17] F. Xu, H. U. Krieger, “Integrating shallow and deep nlp for information extraction”, In *RANLP 2003*.

[18] H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, N. R. Shadbolt, “Automatic ontology-based knowledge extraction from web documents”,*Intelligent Systems, IEEE*, *18*(1), 14-21, 2003.

[19] Wimalasuriya,D. C., Dou, D.,“Ontology-based information extraction: An introduction and a survey of current approaches”, *Journal of Information Science*, *36*(3), 306-323, 2010.

[20] D. Maynard, M. Yankova, A. Kourakis, A. Kokossis, “Ontology-based information extraction for market monitoring and technology watch”, In *ESWC Workshop” End User Apects of the Semantic Web”), Heraklion, Crete, 2005.*

[21] H. Saggion, A. Funk, D. Maynard, K. Bontcheva, “Ontology-based information extraction for business intelligence”, In *The Semantic Web*, pp. 843-856, Springer Berlin Heidelberg, 2007.

[22] I. Augenstein, S. Padó, S. Rudolph, “LODifier: generating linked data from unstructured text”. In *The Semantic Web: Research and Applications*, pp. 210-224, Springer Berlin Heidelberg, 2012.

[23] A. Miller, George, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller. "Introduction to wordnet: An on-line lexical database" International journal of lexicography 3, no. 4, pp. 235-244, 1990.

[24] M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi, E. Fekri, M. Monshizadeh, S. M. Assi. "Semi automatic development of farsnet; the persian wordnet." In *Proceedings of 5th Global WordNet Conference, Mumbai, India*, 2010

[25] C. Ramakrishnan, K. J. Kochut, A. P. Sheth, "A framework for schema-driven relationship discovery from unstructured text", In *The Semantic Web-ISWC 2006*, pp. 583-596, Springer Berlin Heidelberg, 2006.

[26] K. Byrne, E. Klein, "Automatic extraction of archaeological events from text",*Proceedings of Computer Applications and Quantitative Methods in Archaeology, Williamsburg, VA, 2010.*

[27] L. Dali, D. Rusu, B. Fortuna, D. Mladenic, M. Grobelnik, "Question answering based on semantic graphs", In *Proceedings of the workshop on semantic search, 2009.*

[28] A. Yates, O. Etzioni, "Unsupervised Methods for Determining Object and RelationSynonyms on the Web",*Journal of Artificial Intelligence Research*,34, 255-296, 2009.

[29] E. Agichtein, "Scaling Information Extraction to Large Document Collections",*IEEE Data Eng. Bull.*, *28*(4), 3-10, 2005.

[30] N. Nakashole, M. Theobald, G. Weikum, "Scalable knowledge harvesting with high precision and high recall", In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 227-236), ACM, 2011.

[31] A. Sebti and A. Barfroush, "A new word sense similarity measure in wordnet," *Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on*, pp. 369-373, 2008.

[32] M. Montazery and H. Faili, "Automatic Persian wordnet construction." In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 846-850, 2010.

[33] M. Jorge, M. A. Marzal, J. Lloréns, J. Moreiro. "WordNet applications." In *GLOBAL WORDNET CONFERENCE*, vol. 2, pp. 270-278. 2004.

[34] T. Berners-Lee, J. Hendler, O. Lassila, "TheSemantic Web," Scientific American, pp.35-43, 2001.

[35] E. Kuzey, G. Weikum, “Extraction of temporal facts and events from Wikipedia”, In *Proceedings of the 2nd Temporal Web Analytics Workshop*. pp. 25-32, ACM, 2012.

[36] E. Agichtein, L. Gravano,“Snowball: Extracting Relations from Large Plain-Text Collections”,*In Proceedings of the fifth ACM conference on Digital libraries, pp. 85-94, ACM, 2000.*

[37] M. Schuhmacher, S. P. Ponzetto, “Knowledge-based graph document modeling”, In Proceedings of the 7th ACM international conference on Web search and data mining, pp. 543-552, ACM, 2014.

[38] S. Soderland, J. Gilmer, R. Bart, O. Etzioni, D. S. Weld, “ Open Information Extraction to KBP Relations in 3 Hours”, 2014.

[39] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, S. Soderland, “TextRunner: open information extraction on the web”. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 25-26,2007.

[40] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, “Open Information Extraction from the Web”, In *IJCAI*,Vol. 7, pp. 2670-2676, 2007.

[41] M. Banko, O. Etzioni, T. Center, “The Tradeoffs Between Open and Traditional Relation Extraction”, In *ACL*, Vol. 8, pp. 28-36, 2008.

[۴۲] م. شمس‌فرد، ف. جعفری‌نژاد، "استخراج روابط معنایی میان فعل و وابسته‌های آن از متون زبان فارسی"، فصل‌نامه پازند سال ۸ شماره ۳۰ پاییز، ۱۳۹۱.

[۴۳] م. رسولی، ب. م. بیدگلی، ه. فیلی، م. امینیان، "استخراج بی‌ناظر ظرفیت فعل در زبان فارسی"، فصل‌نامه علمی- پژوهشی پردازش علائم و داده‌ها، شماره ۲ پیاپی ۱۸، ۱۳۹۱.

[۴۴] م. شمس‌فرد، ب. صراف‌زاده، "اجتماع نمونه‌دهی هستان‌شناسی و حاشیه‌نویسی معنایی متون فارسی در سیستم POPTA"، نشریه علمی پژوهشی انجمن کامپیوتر ایران، مجلد ۶، شماره ۳ (الف)، ۱۳۸۷.

[45] E. Pianta, L. Bentivogli, and C. Girardi, “MultiWordNet: developing an aligned multilingual database,” *the First International Conference on Global WordNet*, Mysore, India: 2002.

[46] D. Tufi,V. M. Barbu, L. Bozianu, C. Mihil, "Romanian Wordnet: current state, new developments and applications." In *Proceedings of the 3rd Conference of the Global WordNet Association (GWC’06)*, pp. 337-344, 2006.

[47] H. Isahara, F.Bond, K. Uchimoto, M. Utiyama, K. Kanzaki. "Development of the Japanese WordNet", 2008.

[۴۸] پروژه دادگان وابستگی فارسی، گروه پژوهشی دادگان، تابستان ۱۳۹۱.

[49] J. Nilsson, S. Riedel, D. Yuret, “The CoNLL 2007 shared task on dependencyparsing." In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, pp. 915-932, 2007.

پیوست

پیوست- الف

پیکره تجزیه‌ی وابستگی دادگان بر اساس قالب معیار همایش زبان‌شناسی رایانه‌ای و پردازش زبان طبیعی بر روی پیکره‌های وابستگی فراهم شده است. در این پیکره، ویژگی‌های ساخت‌واژه‌ای شامل شخص، شمار، وجه و نمود برای فعل و چسبیدگی واژه در نظر گرفته شده است. چسبیدگی واژه شامل حالتی است که واژه باید منقطع شود و به دو رشته تبدیل شود، مانند «گفتمش» به «گفت» و «ش» [48]. در جدول (الف-۱) اختصارات موجود در برچسب‌های اجزای کلام و ویژگی‌های ساخت‌واژی نشان داده شده است.

جدول الف-۱: اختصارات موجود در برچسب‌های اجزای کلام و ویژگی‌های ساخت‌واژی [48]

اختصار	توضیح	اختصار	توضیح
ACT	معلوم	ADJ	صفت
ADR	نقش‌نمای ندا	ADV	قید
AJCM	صفت تفضیلی	AJP	صفت مطلق
AJSUP	صفت عالی	AMBAJ	صفت مبهم
ANM	جاندار	CONJ	نقش‌نمای همپایگی
CREFX	بازتابی مشترک	DEMAJ	صفت اشاره
DEMON	اشاره	EXAJ	صفت تعجبی
IANM	بی‌جان	IDEN	شاخص
INTG	پرسشی	JOPER	شخصی پیوسته
MOD	وجهی	N	اسم
PART	جزء دستوری	PAS	مجهول
PLUR	جمع	POSADR	نقش‌نمای ندا پسین
POSNUM	صفت شمارشی پسین	POSTP	حرف اضافهٔ پسین
PR	ضمیر	PRADR	نقش‌نمای ندا پیشین
PREM	پیش‌توصیف‌گر	PRENUM	صفت شمارشی پیشین
PREP	حرف اضافهٔ پیشین	PSUS	شبه‌جمله
PUNC	علامت نگارشی	QUAJ	صفت پرسشی
RECPR	ضمیر متقابل	SADV	قید مختص
SEPER	ضمیر شخصی جدا	SING	مفرد
SUBR	نقش‌نمای وابستگی	UCREFX	ضمیر بازتابی غیرمشترک
V	فعل		

در جدول (الف-۲) وجه- نمود- زمان در فعل‌های زبان فارسی نمایش داده شده است.

جدول الف-۲: وجه- نمود- زمان در فعل‌های زبان فارسی [48]

وجه نمود زمان	اختصار	مثال (خوردن)
حال امری	HA	بخور
آیندهٔ اخباری	AY	خواهم خورد
گذشتهٔ نقلی استمراری اخباری	GNES	می‌خورده‌ام
گذشتهٔ بعید استمراری اخباری	GBES	می‌خورده بودم
گذشتهٔ استمراری اخباری	GES	می‌خوردم
گذشتهٔ نقلی اخباری	GN	خورده‌ام
گذشتهٔ بعید اخباری	GB	خورده بودم
حال اخباری	H	می‌خورم
گذشتهٔ سادهٔ اخباری	GS	خوردم
گذشتهٔ بعید استمراری التزامی	GBESE	می‌خورده بوده باشم
گذشتهٔ استمراری التزامی	GESEL	می‌خورده باشم
گذشتهٔ بعید التزامی	GBEL	خورده بوده باشم
حال التزامی	HEL	بخورم
گذشتهٔ التزامی	GEL	خورده باشم

در جدول (الف-۳) روابط وابستگی در پیکره زبان فارسی ارائه شده است.

جدول الف-۳: روابط وابستگی در پیکره زبان فارسی [48]

اختصار	توضیح	اختصار	توضیح
ACL	متمم بندی صفت	ADV	قید
ADVC	متمم قیدی فعل	AJCONJ	همپایهٔ صفت
AJPP	متمم حرف اضافه‌ای صفت	AJUCL	بند افزودهٔ فعل
APOSTMOD	وابستهٔ پسین صفت	APP	بدل
APREMOD	وابستهٔ پیشین صفت	AVCONJ	همپایهٔ قید
COMPPP	حرف اضافهٔ تفضیلی	ENC	فعل‌یار پی‌بستی
LVP	جزء همکرد	MESU	ممیز
MOS	مسند	MOZ	مضافٌ‌الیه
NADV	قید اسم	NCL	بند اسم
NCONJ	همپایهٔ اسم	NE	اسم‌یار
NEZ	متمم نشانهٔ اضافه‌ای صفت	NPOSTMOD	صفت پسین اسم
NPP	حرف اضافهٔ اسم	NPREMOD	صفت پیشین اسم
NPRT	جزء اسمی	NVE	فعل‌یار
OBJ	مفعول	OBJ۲	مفعول دوم
PARCL	بند وصفی	PART	افزودهٔ پرسشی فعل
PCONJ	همپایهٔ حرف اضافه	POSDEP	وابستهٔ پسین
PRD	گزاره	PREDEP	وابستهٔ پیشین
PROG	مستمرساز	PUNC	علامت نگارشی
ROOT	ریشهٔ جمله	SBJ	فاعل
TAM	تمیز	VCL	بند متممی فعل
VCONJ	همپایهٔ فعل	VPP	مفعول حرف اضافه‌ای
VPRT	حرف اضافهٔ فعلی		

Department of Computer Engineering

Ferdowsi University of Mashhad

Web Technology Lab

M.Sc. Thesis

Fact Extraction from Persian Text in RDF Format

By:

Tooba fadaee Tabrizi

Supervisor:

Prof. Mohsen Kahani

February 2015