مجله ی علمی پژوهشی
رایانش نرم و فناوری اطلاعات
جلد۳ شماره ۳ سال ۱۳۹۳

توسعه موتور جستجوی فارسی بر اساس بهبود فرآیند یادگیری آنتالوژی

سیما درویشی[۱]، اسدالله شاه بهرامی[۲]و منوچهر نحوی[۳]
[۱]گروه مهندسی کامپیوتر، دانشکده فنی، دانشگاه گیلان، darvishi@msc.guilan.ac.ir
[۲] گروه مهندسی کامپیوتر، دانشکده فنی، دانشگاه گیلان، shahbahrami@guilan.ac.ir
[۳] گروه مهندسی برق، دانشکده فنی، دانشگاه گیلان، nahvi@guilan.ac.ir

چکیده - موتور جستجو ابزاری است که نیازهای اطلاعاتی کاربران را برطرف می کند. کاربران با استفاده از پرس و جوهای مختلف در موتورهای جستجو تلاش می کنند به نتایج مورد نظر خود دست یابند. برخی از این پرس و جوها با استفاده از کلمات کلیدی مؤثر انجام می شود. اگر موتور جستجویی بتواند تا حدودی روابط حاکم بین کلمات را درک نماید مسلماً نتایج بهتری را برای کاربران استخراج می نماید. اما درک این روابط و بیان منظور کاربر تا حدودی به ساختار پرس و جو در هر زبان بر می گردد. انجام جستجوها در موتور جستجوی فارسی با توجه به قوانین نحوی، تلفظی و املایی آن، کار راحتی نیست. به منظور بهبود و افزایش دقت موتورهای جستجوی فارسی می توان از مفهوم آنتالوژی جهت توصیف کلمات و درک مفاهیم آنها استفاده کرد. در این مقاله با توجه به عملکرد موتور جستجو و آنتالوژی، مدلی جهت بهبود استخراج روابط معنایی از زبان فارسی ارائه می شود. در این مدل از یک مجموعه متنی استاندارد فارسی به نام پیکره بی جن خان استفاده می شود. آزمایشات از طریق الگوهای زبانی و نحوی فارسی بدست آمده بر روی برخی متون این پیکره نشان داد که دقت مدل پیشنهادی حدود ۸۷٪ است.

کلید واژه‌ها- موتور جستجو، آنتالوژی، پایگاه دانش، وب معنایی.

۱. مقدمه

با رشد روزافزون علم و منابع گسترده اطلاعاتی شامل فایل های صوتی، ویدئویی، تصاویر و غیره، وب اطلاعات زیادی را در خود جای داده است که موتورهای جستجوی مختلفی در زبان های متفاوت مانند یاهو[۱]، گوگل[۲]، آلتاویستا[۳] و غیره به کاربران امکان می دهند با ورود کلمات کلیدی پاسخ پرسش خود را بیابند[۱]. موتور جستجو این کلمات را از صفحاتی که قبلاً در بانک های اطلاعاتی خود ذخیره کرده جستجو می کند و با توجه به الگوریتم های موجود رتبه بندی[۴]، صفحات یافت شده را امتیازدهی نموده و آنها را به کاربران نشان می دهد[۲].

مشکل اینجاست که استفاده از کلمات کلیدی برای ماشین قابل فهم نیست و اگر به زبان ماشین نزدیکتر باشد، پاسخ مناسب تری مرتبط با موضوع پرسش یافت می شود[۲]. لذا، ماشین ها باید از مفاهیم بین کلمات مطلع شوند[۳] و استخراج روابط معنایی[۵] و مفهوم اطلاعات موجود در وب برای کامپیوترها قابل فهم گردد.

[۱] Yahoo
[۲] Google
[۳] Altavista
[۴] Ranking
[۵] Extracting Semantic Relations

در وب معنایی از آنتالوژی[۴] برای مفاهیم یک دامنه به طوریکه برای ماشین قابل فهم باشد استفاده می شود. ایجاد دستی[۱] آنتالوژی سخت و دارای احتمال خطاهای انسانی است[۵] زیرا روش های دستی معمولاً نیاز به معماری ساختاریافته[۲] مفهومی توسط خبرگانی دارد که فرهنگ لغت ها[۳] و سایر منابع متنی را بدست آورده و حرفه ای باشند[۳]. لذا، استخراج اطلاعات از منابع به طرز نیمه اتوماتیک[۴] یا تمام اتوماتیک آنتالوژی مورد توجه محققان قرار گرفته است که از اطلاعات وب معنایی است[۶، ۷].

در واقع داده های موجود توسط آنتالوژی پردازش می شوند و مفهوم مرتبط با هر متن از آنتالوژی دامنه بدست می آید و با استخراج اصطلاحات تخصصی هر دامنه نمونه ها ایجاد و در نهایت با توجه به الگوهای موجود، اطلاعات استخراج می شوند. سیستم های موجود برای استخراج آنتالوژی از برخی متون فارسی دارای نقایصی مانند کمبود منابع نحوی جهت بازیابی اطلاعات[۵] هستند[۸، ۹].

در این مقاله پایگاه دانش[۶] موتور جستجوی فارسی به جهت بهینه کردن آن غنی می شود، به این ترتیب که روابط بین واژگان برای یک ماشین قابل فهم شده و در هنگام پرسش کاربران، بتواند نزدیک ترین پاسخ را به آنان برگرداند.

در اینجا از طریق الگوهایی که توسط آنتالوژی بدست می آید، فرهنگ لغتی با گستردگی روابط بین کلمات به وجود می آید به طوریکه در نتیجهٔ جستجوی کاربران، معادل واژه ها و روابط بینشان نیز از طرف ماشین قابل تشخیص باشد. به عنوان مثال اگر کاربری ترکیب «شاخه ِ درخت» را جستجو نماید، موتور جستجو بتواند نوع ترکیب و نوع ارتباط بین کلمات شاخه و درخت را با یکدیگر تشخیص داده و آنها را جهت نمایش صحیح استخراج نماید و از طریق الگوی بدست آمده نوع ارتباط سایر ترکیبات مشابه نیز بازیابی شوند. به جهت این غنی سازی از یک منبع متنی موجود در زبان فارسی به نام پیکره متنی بی جن خان[۷] استفاده می شود[۱۰]. این پیکره دارای متون با موضوعات گوناگون و برچسب های تعریف شده برای کلمات زبان فارسی است. از این رو می توان با توجه به ویژگی های زبان فارسی و استفاده از پیکره متنی مذکور، فرآیند استخراج روابط در متون فارسی را توسط آنتالوژی بهبود بخشید.

در راستای انجام فعالیت های فوق الذکر، ابتدا در بخش ۲، برخی از کارهای مرتبط در زمینهٔ یادگیری آنتالوژی معرفی می شوند. سپس در بخش ۳، اساس روش پیشنهادی به جهت استخراج روابط مفهومی زبان فارسی توضیح داده می شوند. در بخش ۴ ارزیابی سیستم پیشنهادی در راستای توسعهٔ موتور جستجو بیان می شود.

۲. کارهای مرتبط

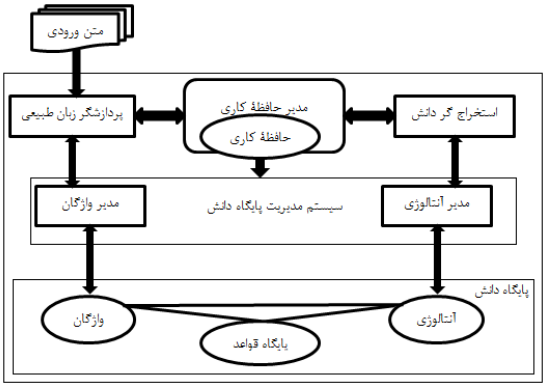
با بررسی های به عمل آمده بر روی تعداد زیادی از مقالات موجود در زمینهٔ یادگیری آنتالوژی روش های مهم آن به دو

[۱] Manual
[۲] Structured Architecture
[۳] Dictionaries
[۴] Semi Automatic
[۵] Information Retrival
[۶] Knowledge Base
[۷] Bijankhan

دستهٔ کلی روش های ساخت آنتالوژی و استخراج روابط تقسیم می شوند:

۱،۲. روش های ساخت آنتالوژی

روش های ساخت آنتالوژی از طریق متون زبان طبیعی و ویکی پدیا[۱] انجام شده است[۱۱، ۱۲]. کار اول برگرفته از دانش مفهومی و ساخت آنتالوژی پایه بر اساس متون زبان طبیعی و ترکیبی از روش های نمادین منطقی، زبانی، الگویی و مکاشفه ای است. آغاز کار با یک هستهٔ کوچک آنتالوژی است که پس از دریافت متون ورودی لایه های آنتالوژی بر روی هستهٔ اولیه ساخته می شوند و بر اساس شرایط قلمرو قابل تغییر است. در این روش، ورودی ها متون زبان فارسی هستند و خروجی، واژگان و آنتالوژی بسط یافته و یاد گرفته شده از متن می باشد. طرحی از سیستم به صورت شکل ۱ است که توضیحات آن در زیر آمده است:

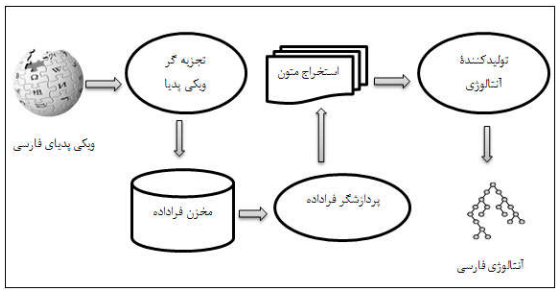


شکل ۱: طرحی از سیستم ساختار یادگیر آنتالوژی

[۱] Wikipedia

با توجه به شکل، چهار بخش اصلی و اجرایی آن به صورت پردازشگر زبان طبیعی، استخراجگر دانش، مدیران پایگاه دانش و حافظهٔ کاری مشخص شده اند. در پایگاه دانش، واژگان و آنتالوژی ها پویا هستند ولی پایگاه قواعد که شامل قوانین و الگوهاست، تغییر ناپذیر است[۱۱، ۲۴].

کار دیگر انجام شده ساخت آنتالوژی با استفاده از ویکی پدیا است که در این روش از یک منبع اطلاعاتی برخط مهم به نام ویکی پدیا استفاده شده است و دلیل آن نیز حالت سلسله مرتبه ای بودن اطلاعات این منبع است که شمای آن را به گراف نزدیک می کند و از این خاصیت آن می توان در ساخت آنتالوژی بهره برد. بین صفحات طبقه بندی شدهٔ ویکی پدیا نوعی ارتباط مفهومی وجود دارد که تحت عنوان "Is-Related-To" از آن یاد می شود و در این روش ساخت آنتالوژی در نظر گرفته شده است.



شکل ۲: فرآیند تولید آنتالوژی فارسی

با توجه به شکل ۲ در تولید آنتالوژی سه بخش اصلی تجزیه گر ویکی پدیا، پردازشگر فراداده و تولیدکنندهٔ آنتالوژی فعالیت دارند که این روش کار خود را بر اساس روش پرسش کابران شروع می کند و اولین فاز سیستم بازیابی اطلاعات محسوب می شود[۱۲، ۱۳].

۲،۲. روش های استخراج روابط

با مطالعات گسترده ای که بر روی روش های استخراج روابط متون انجام شده است، می توان به روش های بسیاری

اشاره کرد که در این قسمت به برخی از آنها اشاره می شود. از جمله این روش ها می توان به بهبود فرآیند یادگیری آنتالوژی از متن به کمک استخراج روابط معنایی از لینک دیتا[۱][۲۰] یاد کرد که به مطالعهٔ معانی پنهان در متن پرداخته است و از طریق یادگیری آنتالوژی از متن، معانی پنهان درون نوشتار را تشخیص داده است. با توجه به فاصله عمیق بین درک انسان و درک ماشین از یک متن، الگوریتم جدیدی برای کشف کلاس های خاصی از دانش در متن ارائه شده که از لینک دیتا در کنار متن خام کمک می گیرد و منجر به استخراج روابط غیر سلسله مراتبی جدید از متن می گردد[۲۰].

همچنین الگوریتم Context Similarity جهت تشخیص تطابق دامنه یک منبع در متن و یک منبع مشابه در لینک دیتا ارائه شده است که داده های متون را از لینک دیتا در وب جمع آوری می کند و یک آنتولوژی چندزبانه تولید می کند که برای غنی سازی آنتالوژی استخراجی از متن استفاده می شود[۲۰].

کار دیگر داده آمیزی بر اساس مدل JDL[۲] است. به طور کلی توسط داده آمیزی می توان اطلاعات را یکپارچه کرد و هدف از این کار تهیه داده های مشخص و قابل فهم مرتبط با موجودیت ها و ارتباط بین آنها است که منجر به استخراج یک دانش جدیدی خواهد شد. از رایج ترین مدل های داده آمیزی، مدل JDL است. این ساختار مشکل چالش معنایی موجود در سیستم های داده آمیزی را با افزودن مفهوم به آنها حل می نماید مانند شناسایی دقیق موجودیت ها و آنتالوژی ها و دستورات نحوی که در هر کدام وجود دارند[۱۴].

کار دیگر، آنتالوژی توسعه یافته WordNet است که دستی ساخته شده و بر روی مقوله های دستوری اسم، صفت، فعل، قید و مجموعه هایی از واژه های مترادف قابل تمایز است[۱، ۱۸]. مجموعه های مترادف از روابط فراواژگان- فرووواژگان سازماندهی می شوند، مانند "نوک و بال ها" که اجزای مفهوم "پرنده" هستند. در همین راستا WordNet فارسی به نام فارس نت[۳] برای زبان فارسی شامل دانش مفهومی و نحوی ساختار افعال انجام شده است[۱۹].

از جمله کارهای دیگر آنتالوژی دالگرن است که به تفاوت نوع ارتباط بین کلمات مانند رابطه های "هست یک" و "تشکیل می شود از" پرداخته است. به عنوان مثال واژه "گله" فرووواژه "حیوان" است ولی رابطه آن " تشکیل می شود از " است.

تنها کار مطرح شده در زبان فارسی آنتالوژی هستی[۴][۱] جهت استخراج دانش مفهومی از متون زبان فارسی و ساخت آنتالوژی از روی آنهاست. در این روش عناصر لغوی شامل کلمات، مفاهیم و عناصر آنتالوژی روابط طبقه ای و غیر طبقه ای از متن استخراج می شوند. از جمله روش های دیگری که در این مقاله بیشتر مد نظر است، روش های مبتنی بر الگوهای زبانی است. مهمترین این الگوها شش الگوی هیرست[۵][۱۵، ۱۶] است که شامل شناسایی روابط بین جفت کلمات در اسناد ساختاریافته و یا نیمه ساختاریافته در پردازش زبان طبیعی است.

---

[۱] Data Linked

[۲] Joint Directors of Laboratories

[۳] FarsNet

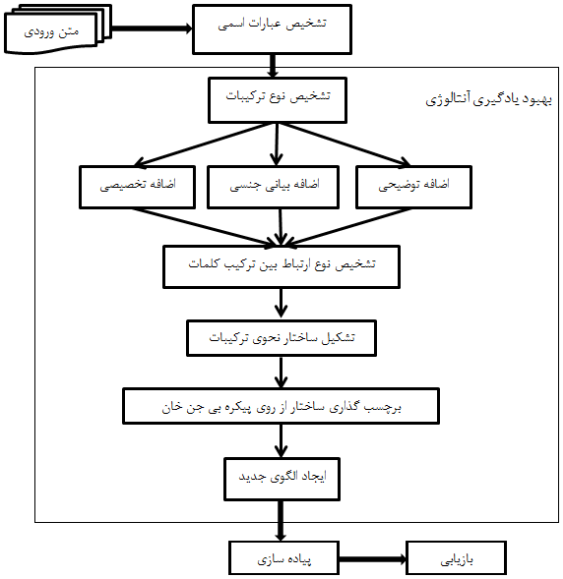
[۴] Hasti

[۵] Hearst

در تمام این کارها در الگوهای عبارات اسمی[1] در زبان فارسی از ترکیبات اضافی فعالیتی انجام نشده است. مسئله اصلی موتور جستجو در بازیابی اطلاعات استخراج اسناد مرتبط با نیاز کاربر از بین سندهای گسترده اطلاعاتی است و نیاز به غنی کردن پایگاه دانش موتور جستجو دارد[۱] که هدف این مقاله است. البته غنی کردن کل پایگاه دانش به دلیل پیچیدگی دستور زبان فارسی امکانپذیر نمی باشد[۱] و در اینجا تنها به بررسی ترکیبات اضافی موجود در عبارات اسمی زبان فارسی و بهبود آن توسط آنتالوژی الگوهای زبانی پرداخته می شود. لذا در این مقاله با استفاده از الگوهای زبانی موجود هیرست برای زبان انگلیسی و تطبیق یافته برای زبان فارسی[۱۷]، الگوهای جدیدی به جهت استخراج روابط مفهومی زبان فارسی بازیابی می شوند.

۳. روش پیشنهادی

مراحل مختلف روش پیشنهادی در شکل ۳ نمایش داده شده است.



شکل ۳: مراحل روش پیشنهادی

در ابتدا برخی از عبارات اسمی دارای ترکیبات اضافی در زبان فارسی مانند اضافهٔ تخصیصی، بیانی جنسی و اضافهٔ توضیحی یا بیانی نوعی را استخراج کرده و سپس رابطهٔ بین این کلمات از طریق نوشتن ساختار ترکیبات مشخص می شوند. از آنجا که تشخیص نوع رابطه بین کلمات در عبارات اسمی کار مشکلی است[۱]، می توان از انواع اسامی مانند عام و خاص بودن آنها و یا در برخی از معرفه و نکره بودن اسامی کمک گرفت که به طبقه بندی در این مرحله نیز تبدیل می شود. سپس نوع رابطه را تشخیص داده و با کمک پیکرهٔ متنی بی جن خان با توجه به هر جزء از کلمه، برچسب گذاری انجام می شود و بعد از برچسب گذاری الگوهای جدید برای هر ترکیب شناسایی می شوند.

در ادامه هریک از بخش های شکل پیشنهادی توضیح داده می شوند.

۱،۳. تشخیص عبارات اسمی در موتور جستجو

[1] Noun Phrase

پرسش کاربر با استفاده از کلمات کلیدی به موتور جستجو داده می شود و تا لحظه دریافت پاسخ به کاربر پردازش هایی در موتور جستجو انجام می شوند. از آنجا که در این مقاله عبارات اسمی از پرسش کاربر در موتور جستجو مد نظر هستند، لازم است به بررسی جایگاه عبارات اسمی و مشکلات و تشخیص آنها در موتور جستجو نیز اشاره شود. به طور کلی موتور جستجو از سه قسمت تشکیل شده است[۲۲]:

خزندهٔ وب[1] که وظیفه جمع آوری مجموعه سندهای موجود را به عهده دارد.

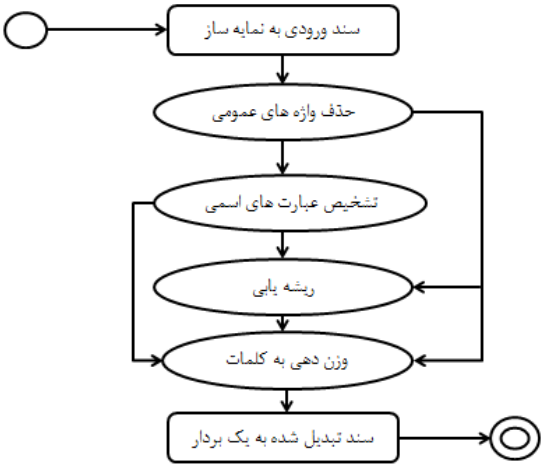
نمایه ساز[2] که مجموعه اسناد جستجو شده توسط جستجوگر را به نماهای قابل استفاده تبدیل می کند.

مدلهای بازیابی اطلاعات و الگوریتم رتبه بندیکه هستهٔ اصلی موتور جستجو از لحاظ رتبه بندی و اولویت در نمایش نتایج جستجو است.

با توجه به استخراج عبارات اسمی مد نظر که در بخش نمایه ساز موتور جستجو مطرح می شود، لازم به معرفی این بخش است.

اطلاعات جمع آوری شده توسط خزنده وب به بخش نمایه ساز فرستاده می شود تا در آنجا مورد تجزیه و تحلیل قرار گیرد. تجزیه و تحلیل می تواند شامل مشخص شدن کلمات کلیدی صفحات، وزن هر یک، اطلاعات آنها و غیره باشد. این بررسی ها که در نمایه ساز صورت می گیرند باید بتوان کلید واژه را از بین تمام اطلاعات یافت که این بر اساس ویژگی های هر زبان و متناسب با قواعد همان زبان، متفاوت است. کلمات کلیدی باید طوری انتخاب شوند که حاوی اطلاعات مورد نیاز برای تشریح سند مورد جستجو باشد. برای استخراج کلمات کلیدی، برخی پیش پردازش ها مثل تعیین کلمات، حذف نقطه گذاری و حذف کلمات کوچکتر از یک آستانه صورت می گیرد. در تعیین کلمات در فارسی ممکن است کلمات چندبخشی هم موجود باشند که باعث استخراج چند کلمه متمایز می شود[۲، ۲۲].

شکل ۴ نمودار نمایه ساز را به همراه کارهای عمومی قابل انجام در این بخش نمایش می دهد[۲۲].



شکل ۴: نمودار نمایه سازی به همراه کارهای عمومی آن

[1] Web Crawler
[2] Indexer

در بخش نمایه سازی فعالیت های زیر انجام می شوند:

حذف واژه های عمومی[۱] که کلمات با تکرار زیاد جهت کاهش حجم پردازش جستجو حذف می شوند.

استخراج عبارات اسمی یعنی کلمات ترکیبی از دو یا چند اسم و نشان دهنده یک عبارت که افزایش دقت بازیابی[۲] را به همراه دارد.

ریشه یابی[۳] در حالت های دستوری کلمات، تفاوت قائل می شود و مستلزم شناخت کافی از دستور زبان است. از کارهای انجام شده در این زمینه می توان به ریشه یاب "حذف وند" بن[۲۱] اشاره کرد که در هر مرحله از بازیابی اطلاعات پیشوندها یا پسوندها را بر می دارد تا کلمه اصلی به دست بیاید. ریشه یابی جهت کاهش حجم نمایه و نیز بهبود بازیابی اطلاعات مورد جستجو استفاده می شود.

وزن دادن به عبارات و واژه ها که از موارد مهم در نمایه سازی است و نقش کلمات را از لحاظ میزان تأثیر آنها به عنوان کلمات کلیدی در متن مشخص می نماید و توسط الگوهای مختلف وزن دهی، جهت تفاوت قائل شدن درکلمات اصلی متن اعمال می شود. در نهایت سند بدست آمده برای ارسال به بخش رتبه بندی و نمایش به کاربر به برداری تبدیل می شود که از نکات الگوریتم های رتبه بندی و مدل های بازیابی محسوب می شود[۲۲].

همانطور که در شکل ۴ مشخص شده است، استخراج عبارات اسمی که در این مقاله مورد بحث است، از فعالیت های موجود در نمایه ساز موتور جستجو است و شامل کلمات ترکیبی از دو یا چند اسم نشان دهنده یک عبارت است و در موتور جستجو باعث افزایش دقت بازیابی کلمات می شود. قبل از معرفی عبارات اسمی به جهت تطبیق ساخت سیستم پردازش و درک متون زبان، لازم است مشکلاتی که در بسیاری از زبان ها به ویژه زبان فارسی وجود دارند به شرح ذیل شناخته شوند:

مشکل کسرۀ اضافی با این توضیح که این علامت در زبان فارسی دارای چند نقش است مانند صفت و موصوف و مضاف و مضاف الیه که در زبان انگلیسی علامت های «of» و «’s» معادل این کسره در بین کلمات است و در همین زبان برای اتصال صفت و موصوف معادلی وجود ندارد[۲۳]. لذا تشخیص این نقش های مختلف توسط ماشین کار مشکلی است و سیستم پردازش نیاز به دانش مفهومی دارد به عنوان مثال حذف کسره اضافی باعث مشکل تشخیص مرز عبارات اسمی می شود.

مشکل پردازش محاسباتی در پردازش متون فارسی که این مشکلات به جهت بدست آوردن الگوی محاسباتی[۴] از زبان برای ماشین به وجود می آیند و شامل کمبود دانش

[۱] Stop Words
[2] Retrival Precision
[۳] Stemming
[4] Computational Pattern

زبانی محاسباتی مانند عدم وجود دستور زبان و الگوها و واژگان[1] و حتی قواعد قابل فهم برای ماشین در زبان هایی مانند فارسی هستند. همچنین عدم وجود ابزارهای پردازش زبان طبیعی برای زبان فارسی مانند تجزیه گرهای نحوی[2] و معنایی کارا و نیز کمبود دانش تخصصی که نیاز به آنتالوژی های مربوطه دارد[۲۳].

در این قسمت به جهت معرفی روش بهبود فرآیند یادگیری آنتالوژی، در زبان فارسی لازم است عبارات اسمی و قوانین آنها بیان شود. علاوه بر گسترش الگوهای موجود[۱۷] در رابطه با استخراج روابط مفهومی بین ترکیبات اضافی در عبارات اسمی، شباهت آن الگوها را برای زبان فارسی استفاده کرده و الگوهای جدیدی نیز با استفاده از دستور زبان فارسی پیشنهاد می شوند. خروجی های این مدل در جهت بازیابی روابط معنایی بین کلمات زبان فارسی می باشند و به همین تناسب درک بهتر موتور جستجوی زبان فارسی از پرسش های کاربران را منجر می شوند.

۲,۳. تشخیص نوع ترکیبات

برخی از عبارات اسمی در زبان فارسی دارای ترکیبات اضافی مانند اضافهٔ تخصیصی، بیانی جنسی و اضافهٔ توضیحی یا بیانی نوعی هستند[۱۷] و به جهت تشخیص نوع ارتباط بین ترکیبات، لازم است در هر عبارت اسمی ابتدا نوع این ترکیبات از لحاظ دستور زبان فارسی مشخص شود.

در هر عبارت اسمی، یک هسته و چند وابسته وجود دارند که وابسته های مد نظر در این قسمت گروه اسمی نام دارند و به هسته اضافه می شوند و ترکیبات اضافی را تشکیل می دهند[۲۴]. مثال هایی در رابطه با انواع ترکیبات اضافی که در اینجا مورد توجه است، در جدول ۱ آمده است:

جدول ۱: مثال هایی از ترکیبات اضافی

نوع ترکیب اضافی	توضیح	مثال
اضافهٔ تخصیصی	مضاف به مضاف الیه اختصاص دارد.	شاخه های درختان
اضافهٔ بیانی جنسی	مضاف جنس مضاف الیه را بیان می کند.	انگشتر طلا
اضافهٔ توضیحی	مضاف توضیحی برای نوع مضاف الیه بیان می کند.	درختان خرما

با توجه به مثال های جدول، رابطهٔ معنایی بین هسته و وابسته که در این ترکیبات اضافی همان مضاف و مضاف الیه است، ممکن است یکی از روابط جزء- کل، مالکیت، جنسیت، نام و نوع باشد[۵]. در زیر برخی ترکیبات اضافی از ادبیات زبان فارسی استخراج شده و مثال هایی نیز از مجموعهٔ پیکرهٔ متنی بی جن خان بیان شده اند:

۱,۲,۳. اضافهٔ تخصیصی

با توجه به تعریف این نوع اضافه، مضاف به مضاف الیه اختصاص دارد، پس می توان اینگونه بیان کرد که کلمه اول را نوشته، بعد از کسره اضافه، کلمه دوم باید به کلمه اول

[1] Terms
[2] Syntactic Parser

اختصاص داشته باشد. به عنوان مثال می توان ترکیب اضافه" شاخهٔ ِ درخت " را نام برد.

۲،۲،۳. اضافهٔ بیانی جنسی

در تعریف این نوع اضافه مضاف، جنس مضاف الیه را بیان می کند و می توان برای آن اینگونه برداشت کرد که کلمه اول را نوشته، بعد از کسره اضافه، کلمه دوم جنس کلمه اول را بیان می کند. مانند انگشتر ِ طلا.

۳،۲،۳. اضافهٔ توضیحی یا بیانی نوعی

همان طور که از تعریف این نوع اضافه مشخص است، مضاف برای تعیین نوع مضاف الیه توضیحی بیان می کند و می توان برای آن اینگونه برداشت کرد که کلمه اول را نوشته، بعد از کسره اضافه، کلمه دوم نوع جنس کلمه اول را بیان می کند. مانند درخت ِ خرما.

در اینجا با تشخیص انواع ترکیبات عبارات اسمی می توان نوع ارتباط بین این ترکیبات را به دست آورد که در ادامه آمده است.

۳،۳. تشخیص نوع ارتباط بین ترکیب کلمات

از آنجا که تشخیص نوع رابطه بین کلمات (شامل آنتالوژی روابط"**has**" و "**part-of**" و "**is-a**" یا "**is-a-part-of**") در عبارات اسمی کار مشکلی است، می توان از انواع اسامی مانند عام و خاص بودن آنها و یا در برخی از معرفه و نکره بودن اسامی کمک گرفت که به طبقه بندی در این مرحله نیز تبدیل می شود[۱] و به مرحله برچسب گذاری متون که در ادامه بیان می شود کمک شایانی می کند. سپس نوع رابطه راحت تر قابل تشخیص می شود. منظور از طبقه بندی همان سلسله مراتب ایجاد ساختار بین ترکیبات است که مرحله بعدی کار محسوب می شود. جهت تشخیص نوع ارتباط بین ترکیب کلمات از مثال های قسمت قبل استفاده شده است.

با توجه به جدول ۱ و مفهوم ترکیبات اضافی، در اضافه تخصیصی وقتی مضاف به مضاف الیه اختصاص دارد، می توان نتیجه گرفت که مضاف قسمتی از مضاف الیه است و رابطه مضاف و مضاف الیه از نوع رابطه "**part-of**" بدست می آید. به عنوان مثال در ترکیب اضافهٔ «شاخه ِ درخت»، شاخه به درخت تعلق دارد.

در اضافه بیانی جنسی نیز مضاف جنس مضاف الیه را بیان می کند به طوریکه در مثال ترکیب اسمی «انگشتر ِ طلا» هم مشخص است، جنس انگشتر طلا است. در نتیجه رابطه مضاف و مضاف الیه از نوع رابطه "**is-a-part-of /is-a**" بدست می آید.

به همین ترتیب در اضافه توضیحی مضاف توضیحی برای نوع مضاف الیه بیان می کند و با توجه به مثال ترکیب اسمی «درختان ِ خرما»، توضیح خرما نوع درختان را مشخص می کند. در نتیجه رابطه مضاف و مضاف الیه از نوع رابطه "**is-a**" بدست می آید.

با توجه به تشخیص ترکیبات اسمی و نوع ارتباط بین آنها می توان در مرحله بعد ساختار ترکیبات را ایجاد کرد.

۴،۳. تشکیل ساختار نحوی ترکیبات

همان طور که پیشتر هم بیان شد، در عبارات اسمی می توان از انواع اسامی مانند عام و خاص بودن آنها و یا در برخی از معرفه و نکره بودن اسامی کمک گرفت که به طبقه بندی و ایجاد ساختار در این بخش نیز تبدیل می شود. در زیر ساختارهای نحوی ترکیبات اضافه تخصیصی، اضافه بیانی جنسی و اضافه توضیحی آمده است.

برای اضافهٔ تخصیصی کلمه اول را نوشته، بعد از کسره اضافه، کلمه دوم باید به کلمه اول اختصاص داشته باشد.

کلمه اول + ِ +کلمه بعد از کسره اضافه به کلمه اول اختصاص دارد، یعنی:

اسم مفرد عام + کسره اضافه + اسم مفرد عام

مثال: شاخهٔ ِ درخت

برای اضافهٔ بیانی جنسی کلمه اول را نوشته، بعد از کسره اضافه، کلمه دوم جنس کلمه اول را بیان می کند.

کلمه اول + ِ + کلمه بعد از کسره اضافه جنس کلمه اول را بیان می کند، یعنی:

اسم مفرد عام معرفه + کسره اضافه + صفت ساده

مثال: انگشترِ طلا

برای اضافهٔ توضیحی یا بیانی نوعی کلمه اول را نوشته، بعد از کسره اضافه، کلمه دوم نوع جنس کلمه اول را بیان می کند.

کلمه اول + ِ + کلمه بعد از کسره اضافه نوع جنس کلمه اول را بیان می کند(حتماً اسم است)، یعنی:

اسم مفرد عام + کسره اضافه + اسم مفرد خاص

مثال: درخت ِ خرما

۵،۳. برچسب گذاری ساختار از روی پیکره بی جن خان

بعد از ایجاد ساختار ترکیبات، با کمک پیکرهٔ متنی بی جن خان با توجه به هر جزء از کلمه، برچسب گذاری انجام می شود که نام برخی از برچسب ها و توضیح آنها در جدول ۲، آورده شده است.

جدول ۲: تعریف برخی از برچسب ها

نام برچسب	توضیح برچسب
ADJ	صفت
ADJ_SIM	صفت ساده
N_SING	اسم مفرد
N_PL	اسم جمع
GEN	کسرهٔ اضافه
...	...

در ابتدا برچسب گذاری برای اضافه تخصیصی که ساختار نحوی آن از مرحله قبل بدست آمده است انجام می شود. برای ترکیب ساختار نحوی "اسم مفرد عام + کسره اضافه + اسم مفرد عام"، هر یک از اجزای ساختار، جزء به جزء برچسب گذاری می شوند و معادل آنها به صورت زیر بدست می آیند:

N_SING_COM_GEN+N_SING_COM

به همین ترتیب در اضافه بیانی جنسی برای ترکیب ساختار نحوی " اسم مفرد عام معرفه + کسره اضافه + صفت ساده " که از مرحله قبل استخراج شد، برچسب گذاری های زیر بدست آمده است:

N_SING_COM_GEN+ADJ_SIM

در اضافه توضیحی نیز برای ساختار نحوی " اسم مفرد عام + کسره اضافه + اسم مفرد خاص "، برچسب های زیر بدست آمده است:

N_SING_COM_GEN+N_SING_PR

۶،۳. ایجاد الگوهای جدید

همان طور که قبلاً نیز گفته شد، الگوهای عبارات اسمی برای بازیابی روابط مفهومی مفاهیم متناظر با صفت و موصوف (مانند ویژگی داشتن) به نام ترکیبات وصفی و یا میان مضاف و مضاف الیه (مانند مالکیت) به نام ترکیبات اضافی است. استخراج نوع رابطهٔ مفهومی در عبارات وصفی از طریق الگوی وصف به راحتی قابل تشخیص است و در این مقاله از آن صرفنظر شده و تنها به بررسی قسمت روابط مفهومی مفاهیم میان مضاف و مضاف الیه پرداخته شده است زیرا به سادگی تشخیص نوع رابطهٔ مفهومی عبارات وصفی نیست[۱۷].

در جدول ۳ مثال هایی در رابطه با ساختار ترکیبات اضافی آمده است:

جدول ۳: مثال هایی از ساختار ترکیبات اضافی

نوع ترکیب اضافی	توضیح	نوع رابطه	مثال
اضافهٔ تخصیصی	مضاف به مضاف الیه اختصاص دارد.	PART-OF/HAS	شاخه های درختان
اضافهٔ بیانی جنسی	مضاف جنس مضاف الیه را بیان می کند.	IS-A	انگشتر طلا
اضافهٔ توضیحی	مضاف توضیحی برای نوع مضاف الیه بیان می کند.	IS-A	درختان خرما

جهت تشخیص رابطهٔ مفهومی میان کلمات به دانش مفهومی در ترکیبات اضافی و موجود در متن نیاز است و در اینجا از روش های مبتنی بر پیکره[۱] های متنی و زبانی استفاده می شود. معروف ترین الگوهای استخراج روابط طبقه بندی شدهٔ عبارات اسمی که پیشتر نیز معرفی شدند، شش الگوی هیرست[۱۶] هستند و در جدول ۴ برای گروه های اسمی NP زبان فارسی تطبیق داده شده اند[۱۷] و مثال هایی نیز در این زمینه آورده شده است:

جدول ۴: مثال هایی از الگوهای تطبیق یافتهٔ هیرست

شمارهٔ الگو	الگو	مثال
۱	<گروه اسمی ش>(مانند | مثل | همچون | چون){<گروه اسمی>،}* [و | یا]<گروه اسمی>	شاعرانی مانند سعدی، حافظ و خیام
۲	<گروه اسمی ش>(منجمله | از جمله){<گروه اسمی>،}* [و | یا]<گروه اسمی>	شاعران از جمله سعدی، حافظ و خیام
۳	<گروه اسمی> {،<گروه اسمی>}* {،} یا <گروه اسمی ش> دیگر	اردک، غاز یا پرندگان دیگر
۴	<گروه اسمی> {،<گروه اسمی>}* {،} و <گروه اسمی ش> دیگر	امیر و حامد و دانش آموزان دیگر
۵	<گروه اسمی ش> {،}شامل {،<گروه اسمی>}* [و] <گروه اسمی>	کشورهای عضو ناتو شامل بلژیک، امریکا و نروژ
۶	<گروه اسمی ش> {،}(به خصوص | مخصوصاً){،<گروه اسمی>}* [و] <گروه اسمی>	همه کشورهای اروپایی، به خصوص انگلستان، فرانسه و آلمان

در این الگوها قاعدهٔ شمول (حرف ش) و نیز کسره در ترکیبات اضافی به چشم می خورند که گاهی ممکن است در صورت درست مشخص نشدن مرز عبارات، استفاده از الگو منجر به خطا شود. در این روابط طبقه بندی شده، استخراج الگوی استثنائات هستی نیز به صورت زیر می باشد:

{هر | همه} <گروه اسمی ش> [بجز | جز] <گروه اسمی> {(و | ،) <گروه اسمی>}*

مثال: همه پرندگان به جز پنگوئن پرواز می کنند.

نوع رابطه، "**is-a**"، است چون پنگوئن یک پرنده است.

۷،۳. پیاده سازی مدل

در این بخش به جهت پیاده سازی نیاز به استخراج الگوهای بیشتر هر یک از ترکیبات و استفاده از الگوهای تطبیق یافته هیرست است. هرچه الگوهای استخراجی بیشتر باشد، اسناد مرتبط تری از متون بازیابی می شوند. لذا با توجه به

[۱] Corpus

ترتیب مراحل قبل و بر اساس دستور زبان فارسی برای اضافه تخصیصی می توان موارد زیر را بدست آورد:

اسم مفرد عام + کسره اضافه + اسم مفرد عام (شاخهٔ ِدرخت)

N_SING_COM_GEN+N_SING_COM

اسم جمع عام + کسره اضافه + اسم جمع عام (شاخه های درختان)

N_PL_COM_GEN+N_PL_COM

اسم جمع عام + کسره اضافه + اسم مفرد عام (شاخه های ِ درخت)

N_PL_COM_GEN+N_SING_COM

و غیره.

همچنین برای اضافه بیانی جنسی موارد زیر استخراج می شوند:

اسم مفرد عام معرفه + کسره اضافه + صفت ساده (انگشترِ طلا - مورد قبول- قطعه صنعتی- طنزسیاه- فیلمنامهٔ کوتاه)

N_SING_COM_GEN+ADJ_SIM

اسم جمع عام+کسره اضافه + صفت ساده (علوم اجتماعی - قوانین اساسی- لوله های پلاستیکی- قطعات صنعتی)

N_PL_COM_GEN +ADJ_SIM

و غیره.

به همین ترتیب برای ترکیب اضافه توضیحی موارد زیر بدست می آیند:

اسم مفرد عام + کسره اضافه + اسم مفرد خاص (درخت ِ خرما)

N_SING_COM_GEN+N_SING_PR

اسم جمع عام + کسره اضافه + اسم مفرد خاص (درختان ِ خرما)

N_PL_COM_GEN+N_SING_PR

اسم مفرد عام + اسم مفرد مکان خاص(شهر ِ تهران)

N_SING_COM_GEN+N_SING_LOC_PR

و غیره.

همان طور که قبلاً بیان شد توضیحات الگوهای تطبیق یافته هیرست برای فارسی نیز مد نظر است که برای هر شش الگو و نیز یک الگوی استثنای هستی موارد زیر استخراج می شوند و با استفاده از روش پیشنهادی، نوع رابطه برای همگی**"is-a"** تشخیص داده شده است و با توجه به جدول ۴ می توان الگوهای زیر را استخراج نمود:

گروه اسمی کلی + {مانند/مثل/همچون/چون} + گروه اسمی که مثال هایی از گروه اسمی اول را با {و/ یا/،} شامل می شود.

مثال: شاعرانی مانند سعدی، حافظ و خیام.

NP_GEN + {مانند/مثل/همچون/چون}+ NP_GEN + PP

NP_GEN+ADV_EXM_GEN+NP_GEN+PP

گروه اسمی کلی + {منجمله/ازجمله } + گروه اسمی که مثال های گروه اسمی اول را با {و/ یا/،} شامل می شود.

مثال: شاعران از جمله سعدی، حافظ و خیام

NP_GEN + { منجمله/ ازجمله} + NP_GEN + PP

NP_GEN+P_GENR+N_SING_COM+NP_GEN+PP

گروه اسمی بیش از دو کلمه که مثال هایی از دسته گروه اسمی دوم است + یا + گروه اسمی کلی + دیگر.

مثال: اردک، غاز یا پرندگان دیگر.

NP_GEN + {یا} + NP_GEN+{دیگر}

NP_GEN+CON_GCOR+NP_GEN+ADJ_SIM

گروه اسمی بیش از دو کلمه که مثال هایی از دسته گروه اسمی کلی دوم است با {و} + گروه اسمی کلی + دیگر

مثال: امیر و حامد و دانش آموزان دیگر

NP_GEN + {و} + NP_GEN+{دیگر}

NP_GEN+ CON_GCOR+ NP_GEN+ ADJ_SIM

گروه اسمی کلی شامل گروه اسمی دوم است + شامل + گروه اسمی معرفی کنندۀ گروه اول که گروه اول را با {و} شامل می شود. مثال: کشورهای عضو ناتو شامل امریکا، بلژیک و نروژ

NP_GEN + {شامل} + NP_GEN+{و}

NP_GEN+ADJ_SIM_GEN+NP_GEN+CON_GCO

گروه اسمی با {هر/همه} + {به خصوص/مخصوصاً} + گروه اسمی که همه گروه اول را شامل می شود به ویژه آنچه در گروه دوم ذکر می شود. مثال: همه کشورهای اروپایی، به خصوص انگلستان، فرانسه و آلمان

NP_GEN + {هر/همه} + {به خصوص/مخصوصاً} + NP_GEN

NP_GEN+QUA_GEN+ADV_NI_NQ_SIM+NP_GEN

استثناء هستی نیز به صورت زیر بدست آمده است:

گروه اسمی با {هر/همه} + {بجز/جز} + گروه اسمی که همه گروه اول را شامل می شود به جز آنچه در گروه دوم ذکر می شود. مثال: همه پرندگان به جز پنگوئن پرواز می کنند.

NP_GEN + {هر/همه} + NP_GEN + {جز/بجز}

NP_GEN+QUA_GEN+ CON_COMP +NP_GEN

در اینجا خلاصۀ تمام الگوهای بدست آمده به صورت جدول ۵ دسته بندی می شوند:

جدول ۵: استخراج الگوهای ترکیبات اضافی

نام ترکیب اضافی	الگو	نوع رابطه
اضافهٔ تخصیصی	N_PL_COM_GEN+N_SING_COM	part-of
اضافهٔ تخصیصی	N_PL_COM_GEN+N_PL_COM	part-of
اضافهٔ تخصیصی	N_SING_COM_GEN+N_SING_CO	part-of
اضافهٔ بیانی جنسی	N_SING_COM_GEN+ADJ_SIM	is-a
اضافهٔ بیانی جنسی	N_PL_COM_GEN+ADJ_SIM	is-a
اضافهٔ توضیحی یا بیانی نوعی	N_SING_COM_GEN+N_SING_PR	is-a
اضافهٔ توضیحی یا بیانی نوعی	N_PL_COM_GEN+N_SING_PR	is-a
اضافهٔ توضیحی یا بیانی نوعی	N_SING_COM_GEN+N_SING_LOC_PR	is-a
هیرست	NP_GEN+ADV_EXM_GEN+NP_GEN+PP	is-a
هیرست	NP_GEN+P_GENR+N_SING_COM+NP_GEN+PP	is-a
هیرست	NP_GEN+CON_GCOR+NP_GEN+ADJ_SIM	is-a
هیرست	NP_GEN+ADJ_SIM_GEN+NP_GEN+CON_GCO	is-a
هیرست	NP_GEN+QUA_GEN+ADV_NI_NQ_SIM+NP_GEN	is-a
هیرست	NP_GEN+QUA_GEN+CON_COMP+NP_GEN	is-a

در این قسمت الگوهای بدست آمده به عنوان ورودی سیستمی که تحت زبان سی شارپ نوشته شده است، به آن داده می شود و باید برای هر الگو در هر ترکیب اضافی عبارت اسمی، نوع ارتباط بین آنها به عنوان خروجی حاصل شود. به جهت تشریح روش پیشنهادی یک مثال در زیر آمده است:

در متنی از حوزه کشاورزی ترکیب اضافهٔ «شاخه ِ درخت» دیده شده است. با توجه به دستور زبان فارسی نوع این ترکیب اضافه تخصیصی است و ساختار آن به شکل زیر است:

اسم مفرد عام + کسره اضافه + اسم مفرد عام

با برچسب گذاری این ترکیب توسط پیکره بی جن خان الگوی زیر حاصل می شود:

N_SING_COM_GEN+N_SING_COM

نوع رابطه **"Part-of"** تشخیص داده شده است چون شاخه قسمتی از درخت است.

الگو و نوع رابطه آن جهت پیاده سازی به برنامه داده می شود. سپس برخی متون انتخابی در دامنه های مورد نظر از پیکرهٔ متنی بی جن خان برای آزمایش به سیستم داده می شوند و در خروجی نوع ارتباط متون استخراج می شوند. در واقع هرگاه متونی در یک دامنه مشخص وارد سیستم شوند، می توانند جهت تعیین نوع رابطه عبارات اسمی از الگوهای بدست آمده استفاده کنند. با استخراج نوع صحیح ارتباط بین ترکیبات، می توان سیستم را مورد ارزیابی قرار داد.

۴. ارزیابی

برای ارزیابی روش پیشنهادی از متون موجود در پیکره متنی فارسی بی جن خان که از موضوعات گوناگون جمع آوری شده است، استفاده می شود. در این قسمت به جهت ارزیابی روش پیشنهادی از مجموعه های متونی موجود در این پیکره، با موضوعات هنری و جغرافیایی انتخاب شده است و نوع عبارات اسمی و نوع رابطه بین ترکیبات و اسناد مرتبط با الگوهایی که از طریق سیستم پیشنهادی بدست آمده است، بازیابی می شوند.

۱،۴. معیار ارزیابی

ارزیابی عملکرد روش پیشنهادی از طریق معیارهای دقت و فراخوانی و پارامتر F[۱] صورت می پذیرد. به این ترتیب که دقت یک سیستم استخراج اطلاعات، نشان دهنده میزان سودمندی اسنادی است که سیستم در پاسخ به پرسش کاربر، بازیابی می کند و منظور از سودمندی اسناد، مرتبط بودن کلمات در متن است که از طریق رابطه (۱) در زیر بدست می آید[۲].

---

[۱] Fmeasure

$$\text{دقت} = \frac{\text{اسناد مرتبط بازیابی شده برای پرس و جو}}{\text{کل اسناد بازیابی شده}} \qquad (۱)$$

در این رابطه از کل اسناد بازیابی شده، سندهای مرتبط بازیابی شده برای پرس و جوی کاربر بدست می آیند.

معیار دیگر ارزیابی، فراخوانی است و نشان دهنده کامل بودن مجموعه سندهای بازیابی شده در پاسخ به پرسش کاربر است و به طور کلی میزان تعیین دربرگیری کلمات در متن را شامل می شود. فراخوانی از رابطه (۲) به دست می آید[۲].

$$\text{فراخوانی} = \frac{\text{اسناد مرتبط بازیابی شده برای پرس و جو}}{\text{کل اسناد مرتبط موجود}} \qquad (۲)$$

در این رابطه از کل اسناد مرتبط موجود در ورودی پرس و جو ها، سندهای مرتبط بازیابی شده بدست می آیند.

به جهت بیشینه کردن دقت بازیابی اطلاعات از پارامتر میانگین هارمونیک دقت و فراخوانی استفاده می شود که از طریق رابطه (۳) بدست می آید.

$$\mathbf{F_{measure}} = \frac{\text{فراخوانی}*\text{دقت}*2}{\text{فراخوانی}+\text{دقت}} \qquad (۳)$$

همچنین برای ارزیابی سیستم از نمودار بررسی دقت استفاده شده است و به همراه نتایج ارزیابی معیارهای فوق در ادامه آمده است.

۲,۴. بررسی نتایج

هر یک از الگوها و نوع روابط استخراجی به برنامه ای که به زبان سی شارپ تهیه شده است، داده می شوند و با انتخاب برخی متون از دامنه های گوناگون موجود در پیکره متنی فارسی بی جن خان، به عنوان ورودی سیستم به آن اعمال می شوند. سیستم با دیدن هر عبارت از متون ورودی مطابق با شباهت آن عبارت به الگو، نوع رابطه بین کلمات را استخراج می نماید. به عنوان مثال اگر هر ترکیب اضافه از نوع بیانی جنسی را ببیند، مانند الگوهای تعریف شده در سیستم عمل می کند و نوع رابطه را به صورت ارتباط **"is-a"** استخراج می نماید.

برای ترکیبات اضافی موجود در عبارات اسمی، الگوها برای انواع گوناگون اسامی از نوع مفرد، جمع، معرفه، نکره، عام و خاص بدست آمده و نوع ارتباط هر یک مشخص شدند. هرچه الگوهای استخراجی بیشتر باشد، اسناد مرتبط تری از متون بازیابی می شوند. لذا با توجه به ترتیب مراحل قبل و بر اساس دستور زبان فارسی می توان الگوهای بیشتری را برای انواع مختلف ترکیبات اضافی بدست آورد. نمونه های بیشتر هر یک از ترکیبات به همراه الگوهای استخراجی و مثال برای آنها در جدول ۵ آمده است.

آزمایشات موجود در این مقاله بر روی دو حوزه متنی از پیکره فارسی بی جن خان انجام شده است که با توجه به عبارات ترکیبی استفاده شده در هر متن، درصد مورد قبولی را به خود اختصاص داده است. همان طور که در جدول ۶ مشخص شده است، دقت سیستم در آزمایش عبارات برای هر دسته از متون به صورت عبارات انتخابی صدتایی با نتیجه مشابه به هم، میانگین حدود ۸۷ درصد بدست آمده است.

- بررسی خطاهای سیستم:

اگر چنانچه نوع رابطه بین ترکیبات عبارات اضافی به درستی مشخص نشوند، الگوهای نادرست از آنها بدست می آیند که سیستم را با خطا مواجه می کند. لذا به دلیل جلوگیری از بوجود آمدن این مشکلات و همچنین به جهت پایداری عملکرد سیستم پیشنهادی، از قواعد موجود در دستور زبان فارسی در تشخیص عبارات استفاده شده است. به جهت تضمین بیشتر در

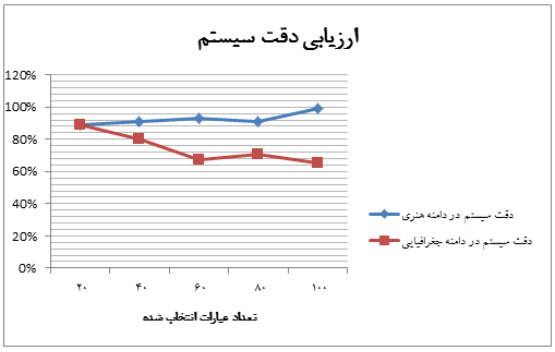
حفظ و نگهداری عملکرد این سیستم از پیکره فارسی استاندارد بی جن خان بهره گرفته شده است.

نتایج ارزیابی سیستم پیشنهادی در جدول ۶ مشخص شده است و به طوریکه دیده می شود از دو دامنه مختلف از متون موجود در پیکره متنی فارسی بی جن خان استفاده شده است.

جدول ۶: ارزیابی نتایج بر اساس دقت و فراخوانی

دامنه متون انتخابی	تعداد عبارات انتخابی	دقت	فراخوانی	$F_{measure}$
هنری	۲۰	٪۸۹	٪۴۰	٪۵۶
	۴۰	٪۹۱	٪۵۰	٪۶۵
	۶۰	٪۹۳	٪۵۹	٪۷۲
	۸۰	٪۹۱	٪۶۲	٪۷۴
	۱۰۰	٪۹۹	٪۶۶	٪۸۰
جغرافیایی	۲۰	٪۸۹	٪۴۰	٪۵۶
	۴۰	٪۸۰	٪۳۰	٪۴۴
	۶۰	٪۶۷	٪۳۰	٪۴۰
	۸۰	٪۷۱	٪۳۰	٪۴۵
	۱۰۰	٪۶۵	٪۲۹	٪۳۴

همان طور که پیشتر بیان شد، در اینجا به جهت بیشینه کردن دقت بازیابی عبارات اسمی از پارامتر F میانگین هارمونیک دقت و فراخوانی، استفاده شده است و نتیجه ارزیابی هر یک از دسته متون انتخابی مشخص شده است. به جهت مقایسه ارزیابی متون در دامنه های مختلف، از لحاظ دقت، نمودار زیر در نظر گرفته شده است.



شکل ۶: نمودار روند دقت سیستم

با توجه به نمودار فوق روند دقت سیستم پیشنهادی برای دسته متون های انتخابی هنری و جغرافیایی، به ازای انتخاب تعداد عبارات مختلف، ارزیابی های متفاوتی دارند و دقت سیستم پیشنهادی به طور میانگین حدود ٪۸۷ می باشد.

۵. نتیجه گیری

در این مقاله معرفی یک سیستم جهت استخراج روابط مفهومی متون زبان فارسی بر اساس آنتالوژی رویکرد مبتنی بر الگو توضیح داده شد و نتایج ارزیابی سیستم توسط معیارهای دقت و فراخوانی صورت پذیرفت.

در انتها ارائه نتایج مقایسه ای با برخی روش ها به جهت بهبود درک روش پیشنهادی به صورت زیر بدست آمده است:

- عدم حذف نشانه گذاری ها در ترکیبات کلمات کلیدی موتور جستجو (با توجه به اینکه در اکثر روش های پیشنهادی به جای استفاده از کسره اضافه، به حذف آن اشاره شده است).
- معنادار شدن بسیاری از ترکیبات دارای کسره اضافه برای ماشین.
- بازیابی عبارات بیشتر از متون و در نتیجه پاسخ نزدیکتری با سرعت بیشتر از طریق پایگاه دانش موتور جستجو برای کاربر.
- استخراج الگوهای بیشتر به جهت بهبود فرآیند سیستم فوق با توجه به دستور زبان نحوی فارسی.
- استخراج روابط مفهومی بین کلمات از طریق الگوهای بدست آمده.

به جهت بهبود فرآیند سیستم فوق می توان الگوهای بیشتری را با توجه به دستور زبان نحوی فارسی استخراج نمود که این کار خود باعث بازیابی عبارات بیشتر از متون شده و در نتیجه

پاسخ نزدیکتری با سرعت بیشتر از طریق پایگاه دانش موتور جستجو به کاربر داده می شود.

مراجع

[۱] فازاول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی"تحلیل نیازمندی های تولید هستان شناسی های عمومی و تخصصی برای زبان فارسی"، زیر پروژه دانشگاه علم و صنعت، ۱۳۸۸.

[۲] فازاول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی"بهینه سازی استفاده از موتورهای جستجو در پیکره های متنی زبان فارسی"، زیر پروژه دانشگاه علم و صنعت، ۱۳۸۸.

[3] Ch. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, L. Prévot, "Ontology and the Lexicon A Natural Language Processing Perspective" , Cambridge University Press 2010.

[4] P. Janardhana, C.Jianhua, "Learning Non-taxonomical Semantic Relations from Domain Texts", Journal of Intelligent Information Systems, Vol.38, No. 1, pp. 191-207, 2012.

[۵] شمس. ف، شمس فرد. م، طالبیان کوچکسرایی. م، "استخراج اتوماتیک اطلاعات بر اساس آنتولوژی"، پایان نامه کارشناسی ارشد، رشته مهندسی کامپیوتر-نرم افزار، دانشگاه علوم و تحقیقات، ۱۳۸۶.

[6] J. P. McCrae, "Automatic Extraction of Logically Consistent Ontologies From Text Corpora", School of Multidisciplinary Sciences,The Graduate Univesity of Advanced Studies, phd thesis, 2009.

[7] N. Kozlova, "Automatic Ontology Extraction for Document Classification", Saarland University, phD thesis, 2005.

[۸] مرتضایی. ل، "مسایل خط و زبا ن فارسی در ذخیره سازی وبازیابی اطلاعات"، فصلنامه اطلاع رسانی، دوره ۱۷، ۱۳۸۰.

[9] H. Fadaei, M. Shamsfard, "Extracting Conceptual Relations from Persian Resources", Seventh International Conference on Information Technology, 2010.

[۱۰] بی جن خان. م، "طرح مدلسازی زبان فارسی"، آزمایشگاه گروه زبان شناسی، دانشکده ادبیات و علوم انسانی، دانشگاه تهران، ۱۳۸۱.

[۱۱] شمس فرد. م، عبدالله زادۀ بارفروش. ا، "ساخت هستان شناسی از روی متون زبان طبیعی"، آزمایشگاه سیستم های هوشمند، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران.

[۱۲] فرهودی. م، محمودی. م، یاری. ع، آزادی نیا. م، "ساخت هستان شناسی فارسی بر اساس ویکی پدیا و کاربرد آن در بسط پرسش"، مرکز تحقیقات هوش مصنوعی ایران.

[13] M. Fahoodi, M. Mahmoudi, A. Mohamma Zare Bidoki, "Query Expansion Using Persian Ontology Derived from Wikipedia", IT Research Faculty, Iran Telecommunication Research Center, World Applied Sciences Journal, IDOSI Publications, 2009.

[۱۴] علیزاده نوقابی. ح، کاهانی. م، بهکمال. ب، شکیبا. ع، "چارچوب داده آمیزی معنایی مبتنی بر مدل JDL"، پنجمین کنفرانس ملی فرماندهی و کنترل ایران، دانشگاه تهران، ۱۳۹۰.

[15] M. Shamsfard, "Lexico-syntactic and Semantic Patterns for Extracting Knowledge from Persian Texts", Faculty of Electrical and Computer Engineering ShahidBeheshti University, Tehran, Iran, International Journal on Computer Science and Engineering(IJCSE), 2010.

[16] M. A. Hearst, "Automatic Acquisitio of Hyponyms from Large Text Corpora", University of California, Fourteenth International Conference on Computational Linguistics, Nantes France, 1992.

[۱۷] شمس فرد. م، عبدالله زاده بارفروش. ا، " استخراج دانش مفهومی از متن با استفاده از الگوهای زبانی و معنایی"، دانشگاه مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تازه های علوم شناختی، سال ۴، شماره ۱، ۱۳۸۱.

[18] Y. Rastegari, H. Abolhassani, B. Zibanezhad, M. Sayadiharikandeh, "Collecting Positive Instances of “instance-of” Relationship in the Persian Language", 2010.

[19] M.shamsfard, A. Hesabi, H.Fadaei, "Semi Automatic Development of FarsNet; The Persian WordNet", computer Engineering, ShahidBeheshti University, Tehran, Iran.

[۲۰] بوشهری. م، زمانی فر. ک، شریعتمداری. ش، "بهبودفرآیند یادگیری آنتولوژی از متن به کمک استخراج روابط معنایی از لینک دیتا ". دومین کنفرانس ملی محاسبات نرم و فناوری اطلاعات، دانشگاه آزاد اسلامی واحد ماهشهر، ۱۳۹۰.

[۲۱] دستور خط: فرهنگستان زبان و ادب فارسی، ۱۳۸۰.

[۲۲] فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی"استخراج نیازمندی های ابزار خطایاب املایی در لایه نحو زبان فارسی به پیکره های فارسی مورد نیاز"، زیر پروژه دانشگاه علم و صنعت، ۱۳۸۸.

[۲۳] فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی"مقدمه ای بر ذخیره و بازیابی اطلاعات متون زبان فارسی"، زیر پروژه دانشگاه علم و صنعت، ۱۳۸۸.

[24] M. Shamsfard, A. Abdolahzadeh Barforoush, " Learning Ontologies From Natural Language Texts", Intelligent Systems Laboratory, Computer Engineering Department, Amir Kabir University of Technology,International Journal of Human-Computer Studies, Elsevier, 2004.