
بسم الله الرحمن الرحيم

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گروه هوش مصنوعی

طراحی و پیاده‌سازی پایگاه داده لغوی فازنت باهدف استفاده در ترجمه ماشینی

دانشجو: محمد نظریان

استاد راهنما:

دکتر مرتضی زاهدی

استاد مشاور:

مهندس مهدی یعقوبی

پایان‌نامه ارشد جهت اخذ درجه کارشناسی ارشد

شهریور ۱۳۹۴

دانشگاه صنعتی شاهرود

دانشکده: مهندسی کامپیوتر و فناوری اطلاعات

گروه: هوش مصنوعی

پایان‌نامه کارشناسی ارشد آقای محمد نظریان

تحت عنوان: طراحی و پیاده‌سازی پایگاه داده لغوی **فازنت**

باهدف استفاده در ترجمه ماشینی

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی: مرتضی زاهدی		نام و نام خانوادگی : مهدی یعقوبی	
نام و نام خانوادگی :		نام و نام خانوادگی :	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی : حمید حسن پور		نام و نام خانوادگی : علی بازقندی	
نام و نام خانوادگی : وحید ابوالقاسمی			
نام و نام خانوادگی :			
نام و نام خانوادگی :			

با استعانت از خدای متعال همان‌که هر دم یادش کنم یادم کند ...

بدین‌وسیله از زحمات تمامی استادان دلسوز و پرتلاش تشکر می‌کنم و برای تمام آن‌ها به‌خصوص استاد زاهدی آرزوی سلامتی و موفقیت در آموزش دانشجویان و کمک به رشد و تعالی جامعه اسلامی را دارم.

از راهنمایی دوست پرتلاشم مهندس امیر کشاورز بهره بردم که ضمن تشکر از ایشان آرزوی موفقیت روزافزون برایش دارم.

تعهد نامه

اینجانب محمد نظریان. دانشجوی دوره کارشناسی ارشد رشته مهندسی کامپیوتر-هوش مصنوعی دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده "پایان‌نامه طراحی و پیاده‌سازی پایگاه داده لغوی FuzzNet باهدف استفاده در ترجمه ماشینی" تحت راهنمایی دکتر مرتضی زاهدی متعهد می‌شوم .

- **تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است 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 یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .**
- **کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نـام « دانشـگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .**
- **حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.**
- **در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافت های آن ها) استفاده شـده اسـت 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 ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .**
- **استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.**

چکیده

با رشد روزافزون علم هوش مصنوعی، روش‌های جدید در این شاخه عملکرد بهتر و خطای کمتر دارند. یکی از مباحث هوش مصنوعی پردازش زبان طبیعی است. مقالات زیادی درباره متن کاوی نوشته شده که بیشتر آن‌ها باهدف تفکیک متون از یک دیگر بوده‌اند و برخی از روش‌های دسته بندی متون از روش‌های فازی استفاده کرده‌اند. یکی از کارهای انجام شده در زمینه پردازش زبان طبیعی و متن کاوی ایجاد پایگاه داده لغوی وردنت است. پایگاه داده لغوی وردنت یک مجموعه از لغات در دسته‌های هم معنی است که بوسیله آن پایگاه دانش ایجاد می‌شود. این پایگاه داده توسط دانشگاه پرینستون ایجاد شد و در زبان‌های دیگر نیز توسعه پیدا کرد. فارس نت به عنوان مدلی از وردنت در زبان فارسی توسط آزمایشگاه پردازش زبان طبیعی در دانشگاه شهید بهشتی پیاده سازی شده است. فارس نت به ازای هر لغت دسته‌های معنایی مختلف را در نظر می‌گیرد و در واقع یک آنتولوژی است. فارس نت مقدمه‌ای بر پایگاه دانش فارسی است.

در این پایان نامه با استفاده از پایگاه داده لغوی فارس نت پایگاه داده لغوی اصلاح شده به نام فازنت ایجاد شده است. در واقع فازنت ترکیبی از فارس نت و مبحث فازی است. با توجه به اینکه در فارس نت ارتباط بین یک لغت با معانی مرتبط آن دقیق است، این ارتباط به جای دقیق بودن، فازی تعریف شده است، یعنی هر لغت با درجات فازی مختلف با مفاهیم خود ارتباط دارد. درجه عضویت یک مفهوم برای یک واژه در صورتی بالاتر است که آن مفهوم به ازای آن لغت بیشتر به ذهن می‌آید. این درجه عضویت فازی در کار ترجمه کمک کننده است بطوریکه آن معنی با درجه عضویت بالاتر برای لغت، انتخاب شده و ترجمه انجام می‌شود.

کلمات کلیدی: فازی، فارس نت، وردنت، ترجمه، پایگاه داده

فهرست مطالب

فهرست اشکال

فهرست جداول

فصل اول

۱- مقدمه

۱-۱-مقدمه‌ای بر پایگاه داده‌های لغوی

خدا برای آخرین پیامبرش معجزه‌ای کامل قرار داد و این معجزه برای پایدار ماندن از قالب گفتار به قالب متن تبدیل شد. متن ابزار انتقال وحی شد و برای همیشه ابزار راهنمای بشریت برای رسیدن به کمال است.

با پیشرفت فن‌آوری پردازش متن برای بهتر کردن متن و اصلاح متن و درک متن بکار گرفته شد.

افزایش روزافزون اطلاعات متنی موجب نیاز بیشتر برای مدیریت متن شد.

مبحث متن‌کاوی[1] زیرمجموعه‌ای از مبحث داده‌کاوی[2] است که در سال‌های اخیر مورد توجه قرار گرفته و مقالات زیادی در این موضوع منتشر شده است؛ که هرکدام با استفاده از ابزارهای مختلف برای یافتن نتایج دلخواه استفاده می‌کنند.

در این پایان‌نامه به دید فازی گونه در متن‌کاوی توجه بیشتری شده است. کاربرد فازی در پردازش زبان طبیعی در بسیاری از پژوهش‌ها مشاهده می‌شود بطوریکه اغلب با خطای کم و محاسبات نسبتاً ساده به نتایج خوبی می‌رسد.

گاهی لغاتی در متن می‌خوانیم که برای ما ناآشناست یا معنی آن را نمی‌دانیم که می‌تواند لغت فارسی یا غیرفارسی باشد. اولین راه‌حل این مشکل جستجوی آن لغت در یک فرهنگ لغت است اما فرهنگ لغت معانی مختلفی برای واژه مورد نظر ارائه می‌دهد که خیلی از آن‌ها با متن ما هم‌خوانی و ارتباط ندارد. مشکل کجاست؟

۱-۲ تعریف مسئله

چرا فرهنگ لغت آن مفهوم و معنی که به دنبال آن هستیم به ما نمی‌دهد؟ پاسخ ساده است. چون فرهنگ لغت از مفهوم و موضوع متن ما آگاهی ندارد و نمی‌داند درباره چه موضوعی باید ترجمه کند.

---

[1] Text-minig
[2] Data-ming

در واقع در فرهنگ لغت ترجمه یا مفهوم یک لغت در موضوعات و متون مختلف به صورت بی نظم و کنار هم ارائه می‌شود البته از لحاظ گرامری یعنی اسم یا فعل یا ... بودن دسته‌بندی می‌شود ولی از حیث مفهوم دسته‌بندی ندارد.

برای حل این مشکل و ایجاد دسته‌بندی مفهومی علاوه بر دسته‌بندی گرامری که باعث ورود بحث هوش مصنوعی به فرهنگ لغت می‌شود پایگاه داده WordNet ارائه شد.

طبقه‌بندی متون به معنی فرآیند دسته‌بندی اسناد با توجه به مجموعه‌ای از یک یا تعداد بیشتری دستهٔ از پیش تعریف شده می‌باشد دستگاه‌های دسته‌بندی خودکار اسناد با افزایش دسترسی به اسناد الکترونیکی و رشد سریع فضای وب، به رویکردی کلیدی به منظور مدیریت اطلاعات و دانش تبدیل گشته‌اند. در دهه‌های اخیر، روش‌های بسیاری به‌منظور دسته -بندی متون پیشنهادشده است که اکثر آنان بر اساس مدل کیسهٔ لغات هستند در این مدل هر عبارت و یا ریشه عبارت، یک مشخصهٔ مستقل در نظر گرفته می‌شود.[۱]

۱-۳ متن‌کاوی

پردازش زبان طبیعی مبحثی مهم در هوش مصنوعی است و دسته‌بندی متون به‌عنوان مسئله‌ای مهم برای پردازش اسناد و داده‌های متنی مختلف با یافتن ساختار گرامری و معنا و نمایش در شکل کاملاً ساختارمند است[2] و[3]

متن‌کاوی فرآیند خودکار استخراج برخی اطلاعات پیش‌تر نامعلوم از تحلیل متن‌های غیر ساختاریافته است.[4] اگرچه متن‌کاوی مشابه داده‌کاوی است اما قدری تفاوت دارند.

داده‌کاوی مربوط به استخراج الگو از پایگاه داده‌ها است در حالی که متن‌کاوی به‌جای پایگاه داده به دنبال زبان طبیعی است. در متن‌کاوی کامپیوتر در تلاش برای استخراج چیزی جدید از منابع موجود و در دسترس می‌باشد. در متن‌کاوی هدف یافتن چیزی است که وجود دارد ولی در داده‌کاوی به دنبال چیزی جدید هستیم.[5]

متن کاوی (TM)[۳] یک زمینه جدید، چالش برانگیز و چند وجهی[۴] است که شامل فضاهایی از دانش مانند علم محاسبات، آمار، پیشگویانه، زبان شناسی و ادراکی است.[6]و[7] متن کاوی در کاربردها و زمینه‌های زیادی به کار گرفته شده است.[8]

۱-۳-۱ تکنیک‌های متن کاوی

تکنیک‌های متن کاوی شامل دسته‌بندی متن، خلاصه سازی، تشخیص موضوع، استخراج مفهوم، تحقیق، سازماندهی متن و غیره است. هر کدام از این تکنیک‌ها می‌تواند در یافتن برخی اطلاعات غیر بدیهی از مجموعه‌ای از اسناد مورد استفاده قرار بگیرد. متن کاوی می‌تواند برای تشخیص موضوع اصلی متن که برای ایجاد رده بندی[۵] از مجموعه متن است مورد استفاده قرار بگیرد.[9]

۱-۳-۲ کاربردهای متن کاوی

متن کاوی در زمینه‌های مختلفی مانند موارد زیر مورد استفاده قرار گرفته است.

- خوشه بندی متن مبتنی بر اسناد وب[10]و[11]و[12]
- دریافت اطلاعات[13]و[14]و[15]
- فشرده سازی و انتقال اطلاعات[16]و[17]و[18]
- ردیابی موضوع[19]و[20]
- خلاصه سازی، سازماندهی، خوشه بندی و ارتباط مفهومی [21]و[17]و[22]و[23]و[12]و[20]و[14]
- نمایش اطلاعات و پاسخگویی به سؤال[24]و[25]
- محتوای عاطفی متون در شبکه‌های اجتماعی برخط[۶] [26]و[27]
- جمع آوری داده‌ها، نمای پایگاه داده، پردازش داده‌ها[25]و[28]و[29]و[30]

---

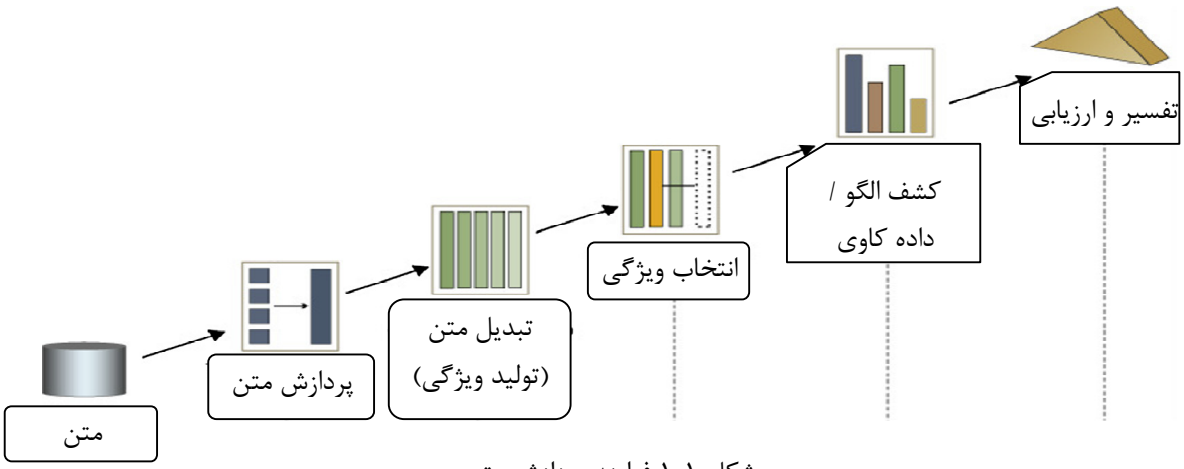
[3] Text mining
[4] Multi-disciplinary
[5] taxonomy
[۵] Online

۱-۳-۳ فرایند متن کاوی

متن کاوی اقدام به تحلیل لغوی[۷] و شناسایی الگو[۸] می‌کند و به مطالعه توزیع تکرار لغت کمک می‌کند. فرایند متن کاوی دارای مراحل اصلی به شرح شکل ۱-۱ است.[9]

۱-۳-٤ خوشه بندی متن

خوشه بندی یکی از مباحث اصلی در ادبیات داده کاوی است. چندین الگوریتم خوشه بندی مختلف برای خوشه بندی متن وجود دارد، ولی بیشتر تکنیک‌های خوشه بندی موجود دچار محدودیت‌های زیادی هستند. روش‌های خوشه بندی موجود با چالش‌هایی نظیر دقت بسیار پایین، زمان دسته‌بندی بالا و غیره مواجه هستند. اخیراً وارد کردن منطق فازی در نتایج خوشه بندی موجب بهتر شدن خوشه بندی شده است. یکی از روش‌های خوشه بندی مبتنی بر فازی پرکاربرد، خوشه بندی C-Meaus (FCM)[۹] می‌باشد.[31]



شکل ۱-۱ فرایند پردازش متن

[7] Lexical analysis
[8] Pattern recognition
[9] Fuzzy C-Means

خوشه بندی سلسله مراتبی اسناد، خوشه‌ها را در درخت یا سلسله مراتبی که مفید برای مرور باشد مرتب می‌کند. ارتباط والد- فرزندی بین گره‌ها در این درخت می‌تواند به‌عنوان ارتباط موضوع- زیر موضوع در نظر گرفته شود.[31]

هر الگو با درجات عضویت فازی مختلف عضو خوشه‌های متفاوت است بطوریکه اگر i شمارنده خوشه باشد آنگاه μ_i درجه عضویت الگو به خوشه شماره i می‌باشد. به عبارت دیگر زوج مرتب (i,μ_i) نشان می‌دهد که الگو با درجه عضویت μ_i به خوشه i تعلق دارد و با در نظر گرفتن حد آستانه برای μ_i ها، خوشه بندی دقیق برای الگو بدست می‌آید.[31]

۱-۳-۵ دسته بندی متن

دسته‌بندی متن اتوماتیک در متن کاوی یک تکنیک شاخص برای مدیریت مجموعه‌های عظیم از اسناد است. با این وجود بیشتر الگوریتم‌های دسته‌بندی متن موجود به‌سادگی توسط اصطلاحات مبهم تحت تأثیر قرار می‌گیرند. لذا توانایی ابهام زدایی یک تفکیک کننده متن به اندازه توانایی صحیح دسته‌بندی کردن آن اهمیت دارد[32]

در مقاله [32] روش دسته‌بندی متن مفهوم گرا مبتنی بر تحلیل مفهومی قاعده‌مند[10] برای آموزش تفکیک کننده با استفاده از مفهوم به‌جای اسناد پیشنهاد شده است تا از ابهامات بکاهد.

رشد اینترنت موجب گسترش فزاینده مقدار متون موجود در پایگاه داده‌ها شده است. به‌منظور سازماندهی، مرور، دریافت و انتشار مؤثرتر چنین داده‌هایی، روش‌های دسته‌بندی اتوماتیک متن توسعه یافته تا اسناد را بر اساس محتوایشان به مجموعه‌ای از دسته‌های از پیش تعریف شده نسبت دهند. یک متن خوب دسته‌بندی شده می‌تواند فیلتر کردن، جستجو و مرور اسناد را هم برای کاربران و هم ابزارهای دریافت اطلاعات تسهیل کند[33].

در حال حاضر دسته‌بندی متن در زمینه‌های مختلف همچون سرویس‌های پزشکی، فیلتر کردن هرزنامه، تشخیص موضوع و دسته‌بندی کتابخانه به‌کارگیری می‌شود[34].

[10] Formal concept analysis

یک تفکیک کننده متن مشابه دیگر تکنیک‌های دسته‌بندی، با استفاده از یادگیری با ناظر توسعه یافته است تا توابع مربوطه را از مجموعه‌ای از اسناد آموزشی با دسته‌بندی دلخواه تولید کند. با این وجود ابهامات، درونی و گنگ بودن متن می‌تواند تفکیک کننده را گمراه کند و نتایج غیرمنتظره تولید شود که این امر دسته‌بندی متن را به‌طور مؤثر متفاوت از دیگر کارهای دسته‌بندی می‌کند[35].

لذا توانایی مدیریت ابهام یکی از مهم‌ترین نشانه‌های یک روش دسته‌بندی متن قوی است[36]و[37].

۱-۴ ساختار پایان نامه، چالش‌ها و فرضیات

در فصل دوم با کارهای تحقیقاتی مرتبط از جمله پایگاه داده لغوی WordNet و فرهنگ طیفی آشنا خواهیم شد. در فصل سوم با ارائه مقدمات تئوری درباره فازی، مبحث متن کاوی و مرورگر فارس نت روش پیشنهادی ارائه می‌شود.

در انتهای فصل سوم روش پیشنهادی یعنی همان فازنت ارائه می‌شود در این بخش فازنت معرفی شده و تفاوت آن با فارس نت و نحوه آموزش آن بررسی می‌گردد.

در فصل چهارم پایگاه داده‌های لغوی WordNet و فارس نت از دید نرم افزاری مورد مطالعه قرار می‌گیرند.

در فصل پنجم الگوریتم فازنت با هدف استفاده در ترجمه ماشینی طراحی، بکار گیری و ارزیابی می‌گردد و در فصل آخر ارزیابی نتایج و کارهای آینده مورد بحث قرار می‌گیرد.

چالش اصلی در ایجاد فازنت یافتن متن ورودی مناسب برای آموزش سیستم فازنت است تا بتواند در کاهش خطا هنگام تست موثر باشد.

در الگوریتم فازنت فرض بر این است که جداسازی متن توسط منطق فازی بدون خطا صورت می‌گیرد.

فصل دوم

۲-کارهای تحقیقاتی مرتبط

Wordnet ۱-۲

۲-۱-۱ معرفی Wordnet

Wordnet یک پایگاه داده واژگانی برای زبان انگلیسی است. wordnet واژه‌های انگلیسی را به مجموعه‌های مترادفی که synsets[11] نامیده می‌شود، گروه‌بندی می‌کند.

WordNetها منابعی الکترونیکی از کلمات یک زبان و روابط بین آن‌ها هستند. اولین و شاید کامل‌ترین wordnet موجود، متعلق به زبان انگلیسی است که ساخت آن از حدود ۱۹۸۰ در دانشگاه Princeton آغاز شد[۱۲]. [38]و[39]

در وردنت اسامی، صفات، افعال و قیود به مجموعه‌ای از لغات مترادف به نام ساین‌ست (*SynSet*) دسته‌بندی‌شده است، بطوریکه هر دسته یک مفهوم مجزا را بیان می‌دارد. وردنت شامل 114648 لغت و 79689 ساین‌ست است و انواع متفاوتی از ارتباطات معنایی در میان ساین‌ست‌ها نگهداری می‌شود. [۱]

۲-۱-۲ تفاوت wordnet با dictionary چیست؟

در واقع پایگاه داده لغوی wordnet ترکیبی از فرهنگ لغت(dictionary) و مفهوم (sense) است. در ظاهر هر دو مشابه هستند که ورودی لغت می‌گیرند و در خروجی توصیفی از آن لغت ارائه می‌کنند. اما در واقع dictionary فرزند WordNet محسوب می‌شود به عبارت دیگر WordNet ترکیبی از dictionary و مفهوم (sense) است؛ یعنی wordnet به ازای لغت ورودی می‌گوید که این لغت معادل چه واژه‌ها و چه مفاهیم دیگری است.

[11] Synonym sets

[۱۲] این WordNet را می‌توانید در http://wordnet.princeton.edu مشاهده نمایید.

در یک WordNet لغات در مجموعه‌هایی سازمان‌دهی می‌شوند که synset[۱۳] نام دارند. اعضای هر synset دارای ارتباط معنایی با یکدیگر بوده و همچنین نقش‌های یکسانی در زبان دارند به‌گونه‌ای که می‌توان آن‌ها را در یک متن به‌جای یکدیگر استفاده نمود. در بین synsetها نیز ارتباطاتی برقرار می‌شود. این ارتباطات می‌تواند هم‌معنایی، تضاد معنایی یا روابطی همچون اسم و صفت و غیره باشد.

نرم‌افزار browser WordNet(ارائه شده توسط دانشگاه Princton) یک واسط کاربری گرافیکی برای نمایش پایگاه داده لغوی WordNet است که بر اساس یافتن sense برای هر لغت با در نظر گرفتن نقش آن در جمله است.

و برای یک لغت با در نظر گرفتن هر مقوله نحوی[۱۴] آن در جمله یعنی اسم، فعل، صفت یا ... sense های آن مقوله را توصیف می‌کند.

مثال:

همان‌طور که در شکل ۲-۱ مشاهده شده است با جستجوی لغت «balloon» در پایگاه داده لغوی WordNet برای هرکدام از مقوله‌های اسم و فعل دو مورد sense بدست آمده است که هرکدام با شماره‌های ۱ و ۲ توصیف شده است. در توصیف sense های مقوله فعل برای هرکدام از sense ها مثالی نیز آورده شده است که این مثال‌ها نشان دهنده پربار بودن پایگاه داده لغوی WordNet است.

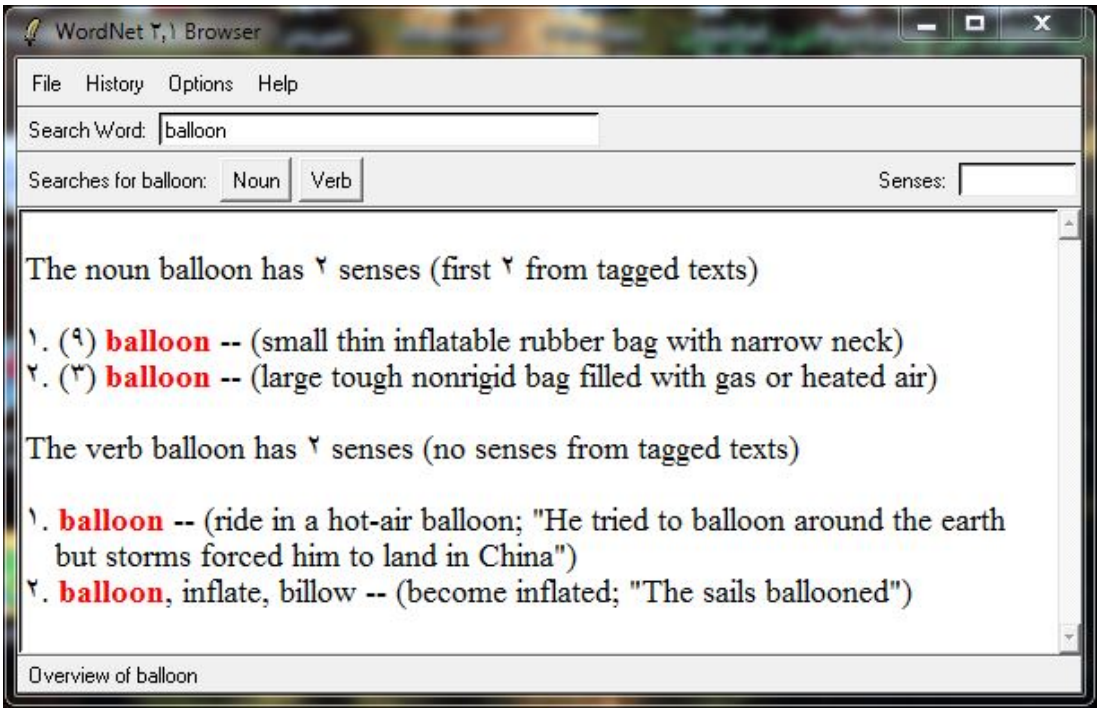
در مثالی دیگر با جستجوی واژه dog در نرم‌افزار WordNet 2.1 browser نتیجه به صورت شکل ۲-۲ است.

همان‌طور که در شکل ۲-۲ مشخص شده است برای واژه «dog» هفت مورد sense مختلف برای مقوله اسم یافت شده است و تنها یک مورد sense برای مقوله فعل دارد.

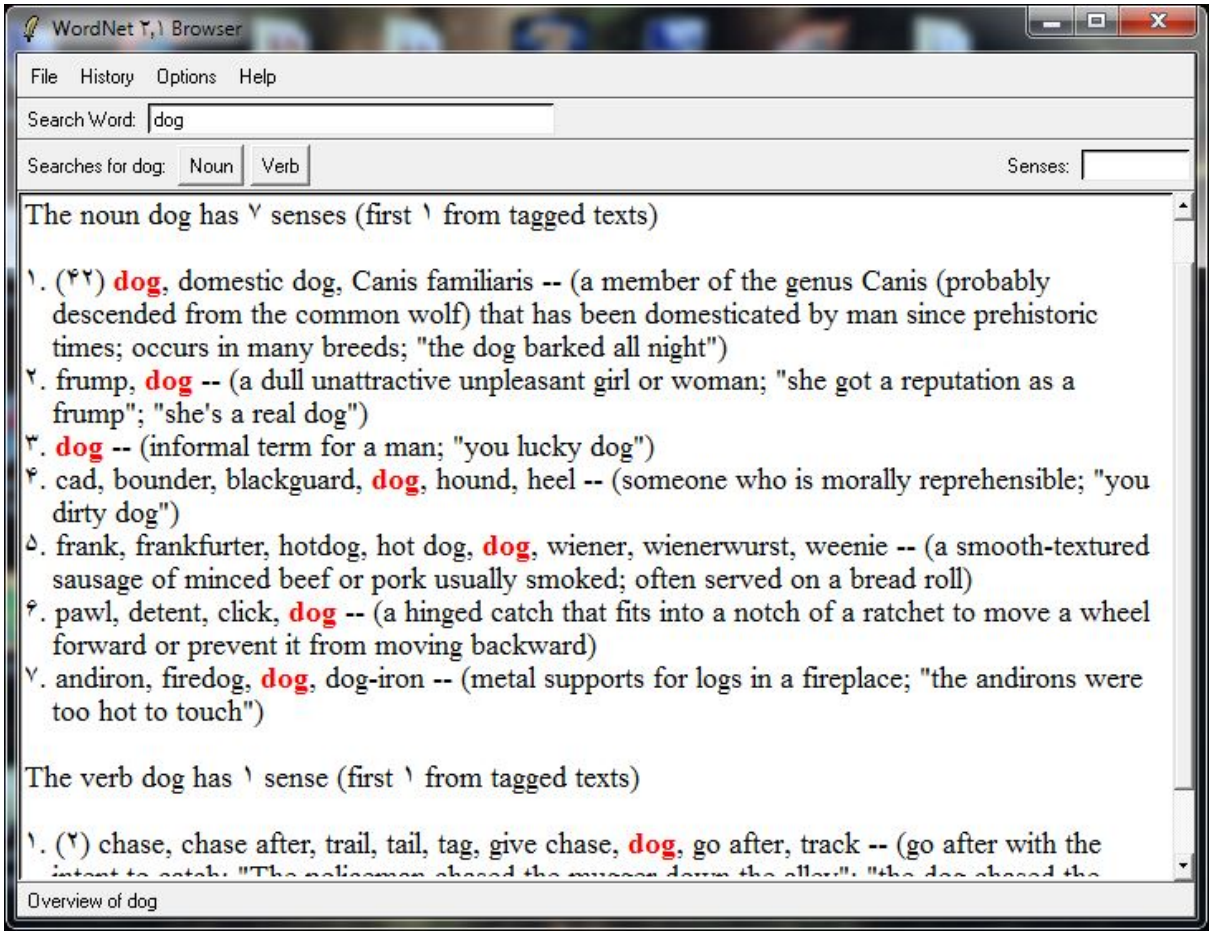
---

[13] Each meaningful concept in WordNet, possibly described by multiple words or word phrase, is called a "synset."

[14] Part of speech

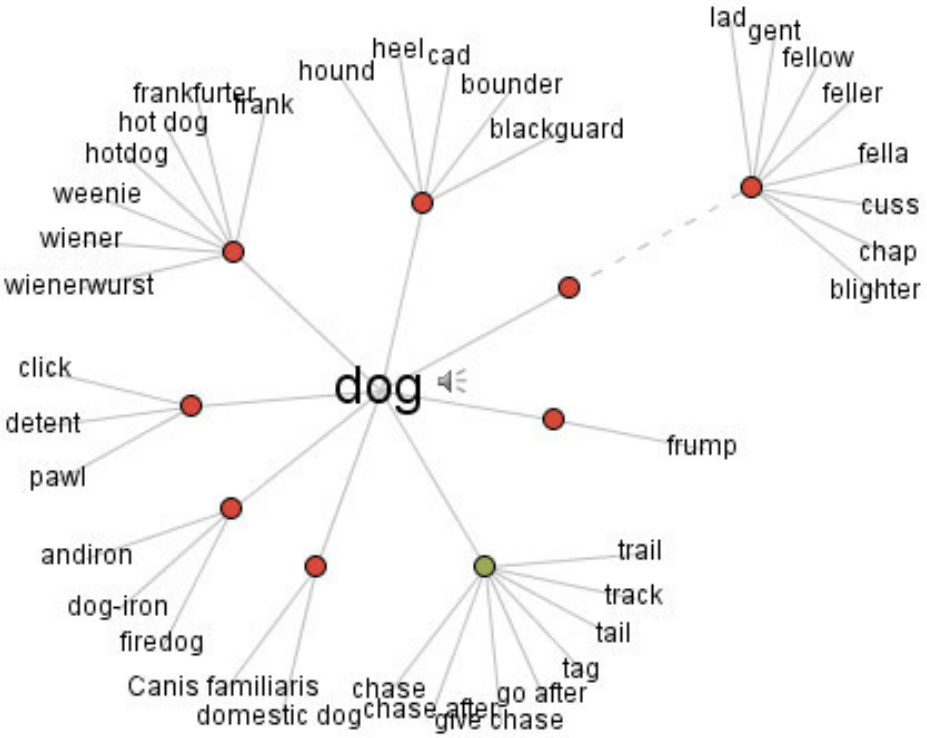


شکل ۲-۱ مثالی از عملکرد WordNet Browser برای واژه balloon



شکل ۲-۲ مثالی از عملکرد WordNet Browser برای واژه dog

در شکل ۲-۳ نمایش دسته‌بندی معنایی به شکل گرافیکی نشان داده شده است.



شکل ۲-۳ نمایش دسته‌بندی معنایی برای واژه dog

۲-۱-۳ کاربردهای WordNet

ساختارهای WordNet می‌توانند نقش مهمی در بازیابی اطلاعات، ترجمه یا تحلیل متون بازی کنند؛ بنابراین تشکیل WordNet برای هر زبانی در تحقیقات مبتنی بر متن مربوط به آن زبان، دارای اهمیت بالایی است. در مورد زبان فارسی تلاش‌های معدودی برای توسعه WordNet صورت گرفته است.

Wordnet یک پایگاه داده واژگانی برای زبان انگلیسی است که اولین بار توسط جورج میلر(George A. Miller[15]) استاد دانشگاه پرینستون[16] و برای زبان انگلیسی ارائه شد. ولی به مرور برای دیگر زبان‌ها ازجمله فارسی هم ایجاد شد.

۲-۱-٤ تفاوت بین wordnet و WordNet:

تفاوت این دو واژه از دید فرهنگ لغت به صورت زیر است:

wordnet به هر پایگاه داده لغوی قابل خواندن برای ماشین بعد از ظهور Wordnet دانشگاه پرینستون[17] گفته می‌شود اما WordNet یک پایگاه داده لغوی قابل خواندن توسط ماشین که به وسیله معانی، سازمان‌دهی و در دانشگاه پرینستون تولید شده است.[18] در واقع تفاوت دو واژه مربوط به عام و خاص بودن آن‌هاست.[19]

پایگاه داده لغوی WordNet واژه‌های انگلیسی را به مجموعه‌های مترادفی که synsets نامیده می‌شود، گروه‌بندی می‌کند.

۲-۱-۵ WordNet به‌عنوان یک ontology

در wordnet واژه‌ها از نظر مفهوم شناسایی و بررسی می‌شوند و در ساختار سلسله مراتبی و ارتباط با واژه‌های دیگر قرار می‌گیرند. ارتباطاتی مانند a part of، a kind of، مترادف، متضاد و ... مفهومی هستند و می‌توان از آن‌ها با عنوان ontology لغوی یادکرد. این ontology قابل استفاده برای ایجاد

---

[15] http://wordnet.princeton.edu/wordnet/about-wordnet/george-miller/

[16] Princeton university

[17] .(any of the machine-readable lexical databases modeled after the Princeton WordNet)

[18] Princeton WordNet (a machine-readable lexical database organized by meanings; developed at Princeton University)

[19] .(همانند Internet و internet)

یک پایگاه دانش (knowledge base) است که البته برای wordnet پایگاه دانش WebKB ایجاد شده است[۲۰].[40] و[41]

۲-۲ آنتولوژی[۲۱]

آنتولوژی ریشه در فلسفه دارد و مبدأ آن را ارسطو می‌دانند.

در فلسفه آنتولوژی شاخه‌ای از علم است که به بررسی موجودات و روابط آن‌ها می‌پردازد. مفهوم آنتولوژی در وب معنایی کمی متفاوت از فلسفه است. آنتولوژی از دو واژه‌ی Onto به معنای هستی و Logia به معنای مطالعه به وجود آمده است و در کل معنی هستی شناسی دارد. آنتولوژی در وب معنایی واژه‌ها و ارتباط بین آن‌ها در دامنه‌ای که استفاده می‌شود را نشان می‌دهد. عناصر اصلی تشکیل دهنده آنتولوژی عبارت‌اند از: مفاهیم، ارتباط بین آن‌ها و خصوصیات آن‌ها. در شکل نقش آنتولوژی در وب معنایی نشان داده شده است.

۲-۲-۱ تعریف آنتولوژی

آنتولوژی را خیلی کوتاه می‌توان به این صورت تعریف کرد: «**آنتولوژی مشخص کردن و تعریف یک مفهوم سازی است.**»

در حالی که کلمات مشخص کردن و تعریف مفهوم سازی باعث بحث‌های زیادی شده‌اند، نکته‌ی اساسی در مورد این تعریف از آنتولوژی موارد زیر هستند:

- یک آنتولوژی مفاهیم، ارتباطات و سایر مختصاتی که برای مدل سازی یک دامنه مورد نیاز هستند را تعریف می‌کنند.

[۲۰] این پایگاه دانش در سایت http://www.webkb.org قابل دسترس می‌باشد.

[21] Ontology

 تعریف یک شکل از تعاریف یک فرهنگ لغات نمایش (کلاس‌ها، ارتباطات و ...) را در بردارد که معانی را برای فرهنگ لغات و قیود رسمی برای استفاده همیشه از آن فراهم می‌کند. یک آنتولوژی لغات و مفاهیمی (معانی) که در تعریف و نمایش یک محظحدوده‌ای از دانش به کار می‌روند را تعیین می‌کنند و بنابراین معانی را استاندارد می‌کنند.

آنتولوژی ها توسط مردم، پایگاه‌های داده و برنامه‌های کاربردی که نیاز به اشتراک گذاری اطلاعات یک دامنه‌ی خاص دارند استفاده می‌شود. در زمینه وب آنتولوژی ها یک فهم مشترک از یک دامنه را تأمین می‌کنند. چنین فهم مشترکی برای حل مشکل چند معنایی لازم است؛ زیرا دو برنامه‌ی کاربردی ممکن است از دو ترم متفاوت برای یک معنای واحد استفاده کنند و یا بالعکس از یک ترم واحد برای دو مفهوم متفاوت استفاده کنند.

۲-۲-۲ عناصر مختلف آنتولوژی

عناصر مختلف آنتولوژی شامل موارد زیر است:

- نمونه‌ها: اشیاء یا نمونه‌ها (اشیای ابتدایی)
- کلاس‌ها: مجموعه‌ها، مفاهیم، انواع اشیاء یا انواع چیزها
- خاصیت‌ها: جنبه‌ها، ویژگی‌ها، قابلیت‌ها، خصوصیات یا پارامترهایی که آن اشیاء (کلاس‌ها) می‌توانند داشته باشند.
- ارتباطات: ارتباطاتی که کلاس‌ها و نمونه‌ها می‌توانند با یکدیگر داشته باشند.
- جملات تابعی: ساختارهای پیچیده‌ای که از یک ارتباط مشخص شکل می‌گیرند که می‌توانند به جای یک جمله یا کلمه‌ی خاص یک عبارت مورد استفاده قرار گیرند.
- قیدها: توضیحاتی که به صورت رسمی بیان می‌شوند تا مشخص کنند که چه چیزی باید صحیح باشد تا اینکه یک حکم به عنوان ورودی مورد پذیرش قرار گیرد.

- قوانین: جملاتی که به صورت اگر- آنگاه استنتاج‌های منطقی که می‌توانند از یک حکم به صورت خاصی به دست آیند را توصیف می‌کنند.
- قواعد کلی: احکام و قوانینی که در یک شکل منطقی با هم یک تئوری را که آنتولوژی در دامنه‌ی یک کاربرد شرح می‌دهد تشکیل می‌دهند.
- رخدادها: تغییر ارتباطات یا خاصیت‌ها

آنتولوژی ها معمولاً در یک زبان آنتولوژی نمایش داده می‌شوند.[۴۲]

آنتولوژی ارتباط بین مفاهیم در اسناد وب و دنیای واقعی را مشخص می‌کند که با این کار اسناد مربوطه توسط ماشین‌ها قابل پردازش و فهم می‌شوند و اشتراک گذاری بین عامل‌ها را تسهیل می‌نماید.

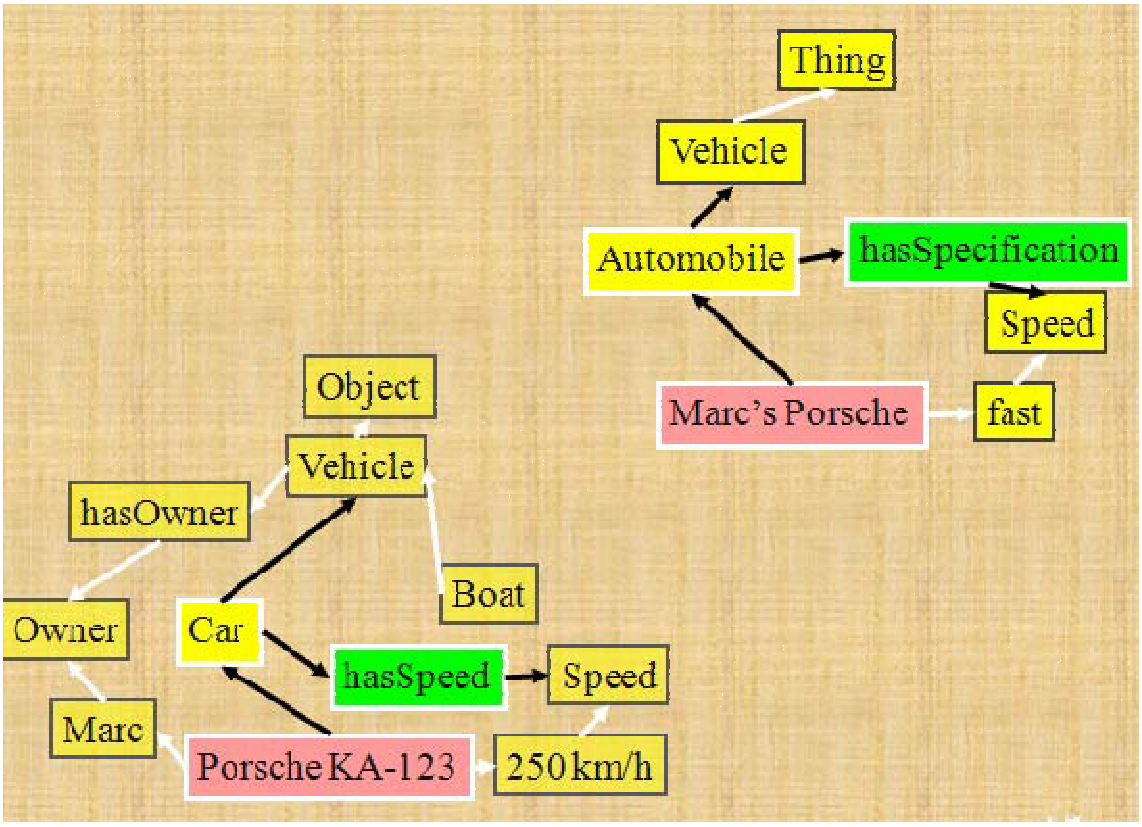
در واقع می‌توان گفت:

Vocabulary+Structure=Taxonomy

Taxonomy+Relationship و Constraints&Rules=Ontology

Ontology+Instance=Knowledge

در فرهنگ وب، آنتولوژی ها مفهومی مشترک از یک دامنه را تأمین می‌کنند یا حداقل فرض می‌شود که این عمل را انجام می‌دهند.[۴۳]



شکل ۲-۴ نمونه‌ای از دو آنتولوژی مربوط به اتومبیل (دوراندیش ۱۳۸۹)

۲-۳ فرهنگ طیفی

کاربرد اصلی فرهنگ طیفی (تزاروس فارسی)، یافتن لغاتی است که شخص به یاد نمی‌آورد (به زبان خودمانی، نوک زبانش است)، که با به یادآوردن و کمک گرفتن از لغات مترادف یا مرتبط با آن انجام می‌گیرد. بهتر است برای روشن شدن موضوع، شیوه‌ی استفاده از این کتاب را با نحوه و هدف استفاده از یک فرهنگ لغت متداول مثل فرهنگ لغات عمید مقایسه کنیم. هنگامی که با لغت ناشناخته‌ای در یک کتاب یا در محاوره برخورد می‌کنید، به یکی از فرهنگ‌ها، مثلاً فرهنگ عمید مراجعه می‌کنید. در چنین فرهنگی، لغات به ترتیب الفبایی مرتب شده‌اند و شما با دانستن یا حدس زدن هجی کلمه‌ی مورد نظر، آن را به سرعت پیدا و معنایش را استخراج می‌کنید.

۲-۳-۱ کاربرد فرهنگ طیفی

تزاروس فارسی یا فرهنگ طیفی یا مقوله‌ای باهدف برعکس استفاده می‌شود؛ یعنی در وضعیتی که شما دنبال لغتی هستید که آن را از یاد برده اما معنای دقیق و مورد استفاده‌ی معمول آن را می‌دانید. یا ایده یا معنایی در ذهن داشته و می‌خواهید بیان کنید اما واژه‌ی مناسب و راضی کننده‌ای برای آن به فکرتان خطور نمی‌کند. در چنین موقعیتی به مجموعه یا فهرستی از لغات نیاز دارید که دارای لغت مورد نظر باشد یا حداقل آن را به یاد شما بیندازد. اینجاست که می‌توانید به فرهنگ طیفی رجوع کنید و با در نظر داشتن یک لغت مشابه یا مرتبط با لغت مورد جستجو (که از این پس آن را لغت آغازین یا سرنخ می‌نامیم) و با استفاده از ایندکس کتاب (نمایه)، شماره‌ی مقوله و نام زیر مقوله‌ی آن را پیدا کنید. برای آغاز جست‌وجو اصلاً لازم نیست از لغت مترادف و نزدیکی شروع کنید بلکه کافی است لغت آغازین (سرنخ) با ایده مورد نظر ارتباطی (تداعی) داشته باشد. به عنوان مثال هنگامی که دنبال بیان ایده‌ی مورد نظر بوده‌اید، حتماً لغات زیادی را به علت عدم کفایت معنایی و نرساندن مقصود کنار گذاشته‌اید؛ کافی است با یکی از همین لغات شروع کنید (فراموش نکنید که هدف شما یافتن لغت یا لغاتی است که مقصود شما را کاملاً ادا کند.) و سپس با در دست داشتن این لغت آغازین (سرنخ) به نمایه (ایندکس) کتاب که به ترتیب الفبایی است رجوع کنید. در مقابل آن لغت سرنخ چند سرفصل و شماره می‌بینید. فرض کنیم دنبال لغت پیشه هستید و لغت اولیه‌ی کار به ذهنتان آمده که گویا نیست. در نمایه‌ی الفبایی لغت کار را پیدا و در مقابل آن سرفصل‌های زیر را مشاهده می‌کنید:

کار

ماموریت ۶۷۲

شغل ۶۲۲

کار ۶۸۲

مشخص است که سرفصل شغل به معنای مورد نظر شما نزدیک‌تر است. پس با توجه به شماره‌ی مقوله‌ی اصلی (رقم ۶۲۲ در جلوی شغل)، به آن مقوله یعنی «۶۲۲ اشتغال» رجوع می‌کنید. مدخل (سرفصل با زیر مقوله) شغل را می‌یابید و همه لغات مترادف و مرتبط را می‌خوانید و یکی را که دقیقاً منظور را می‌رساند، انتخاب می‌کنید؛ یعنی به لغت مطلوب خود می‌رسید.

۶۲۲ اشتغال

...

<< شغل، پیشه، کار،...، فعالیت ۶۷۸، اشتغال، استخدام، خدمت... >>

...

(در این مثال اگر دنبال لغت وظیفه بودید، احتمالاً به سرفصل مأموریت ۶۷۲ مراجعه می‌کردید و برای نوبت‌کاری به سرفصل کار ۶۸۲)

۲-۳-۲ ویژگی فرهنگ طیفی

یک ویژگی عمده‌ی این کتاب، درج عطف به مقولات مرتبط و مشابه در هر مقوله است. (در مثال فوق پس از لغت کار، لغت فعالیت و شماره‌ی ۶۷۸ را می‌بینید). مثلاً لغات گوارایی و سلامت (تندرستی) در مقولات جداگانه‌ای آمده‌اند (گوارایی ۶۵۲ و سلامت ۶۵۰)، اما شماره‌ی مقولهٔ هر یک به‌عنوان عطف در کنار دیگری درج شده است، زیرا

۱) گاهی اوقات و در جملات خاصی می‌توان یکی را به‌جای دیگری استفاده کرد و

۲) احتمال دارد شخصی که دنبال یکی از این دو لغت می‌گردد، لغت دیگری به ذهنش خطور کرده باشد؛ و هدف این کتاب تسهیل دریافتن لغات است.

همان‌طور که گفته شد، تمام لغات متداول فارسی حتی تعدادی از اسامی جانوران و گیاهان و غذاها و غیره (به‌استثنای اسامی اشخاص و اسامی جغرافیایی فاقد تداعی معنایی) در این کتاب آمده است؛ و این کتابِ مرجع، مجموعه‌ی فشرده و کم‌حجمی از واژه‌های فارسی است. مضافاً هر لغت فارسی بسته

به تعداد معانی مختلف آن، در مقولات مختلفی تکرار شده است. به‌عنوان مثال لغت شیر در مقولات: حیوانات، غذا و رطوبت آمده است.

شیر

شیر [لبنیات] ۳۰۱ (مدخل لبنیات در مقوله‌ی ۳۰۱ غذا: خوردن و نوشیدن)

شیرآبیاری ۳۴۱ (در مقوله‌ی ۳۴۱ رطوبت)

پستانداران ۳۶۵ (در مقوله‌ی ۳۶۵ حیوانات)

سرفصل‌ها یا زیرمقوله‌های کتاب در هر مقوله عبارت‌اند از یک یا چند اسم، اسم مصدر، صفت، مضاف‌الیه، مصدر، اصطلاحات امری، ... و قیدهای مرتبط با مقوله‌ی اصلی که در مقابل آن‌ها لغات دیگر به‌صورت زنجیره‌وار درج شده‌اند. در نمایه (ایندکس) کتاب در مقابل هر لغت تعدادی مدخل (سرفصل) و در مقابل هر مدخل، شماره‌ی مقوله درج شده است.

نمایهٔ الفبایی کتاب به‌منظور یافتن مقولات مرتبط به یک لغتِ سرنخ تهیه شده است اما شامل همهٔ لغات زبان فارسی نیست. اصولاً اگر لغتی را در نمایه نمی‌بینید لغت عام‌تر از آن را در نمایه یا اینکه مستقیماً در متن مقوله‌ای جستجو کنید؛ مثلاً اگر لغت سیم‌کارت در ایندکس نیست به موبایل یا تلفن رجوع کنید.

اسامی خاص به‌طورکلی در ایندکس نیست اگرچه در بخش مقوله‌ای کتاب بعضاً وجود دارد. به عنوان مثال نام حیوانات، پرندگان، غذاها (من‌جمله میوه‌ها)، بیماری‌ها، گیاهان (من‌جمله گل‌ها)، سیاره‌ها، عناصر و مواد شیمیایی، داروها، اعداد، لغات کم استفاده و نامأنوس، لباس‌ها، روزهای هفته و نام ماه‌ها، آلات موسیقی، انواع پارچه‌ها، اعضای بدن، اسامی مشاغل، لوازم آرایش و مشابه آن.

لذا برای یافتن چنین لغاتی به مقولات زیر یا لغات عام‌تر نظیر گیاهان رجوع شود:

ظرف ۱۹۴، حیوانات ۳۶۵ ، گیاهان ۳۶۶ ، درخت ۳۶۶، گُل ۳۶۶

بدن انسان ۳۷۱، بیماری و ضعف جسمانی ۶۵۱

غذا: خوردن و نوشیدن ۳۰۱، پوشاک و پوشیدن ۲۲۸

اجزای خودرو ۲۷۴، ردیف‌های موسیقی ◄ گام [موسیقی] ۴۱۰

به ریشه اصلی نیز رجوع کنید مثلاً بی‌شکل در شکل.

چون مقولات متضاد پشت سرهم در کتاب آمده‌اند، اگر در نمایه فقط لغت متضاد واژهٔ موردنظر خود را یافتید از همان عطف استفاده کنید و در مقولات قبلی یا بعدی در متن اصلی کتاب لغت موردنظر خود را بیابید. ضمن اینکه در خود نمایه هم بعضاً لغات متضاد لغت اصلی در همان‌جا عطف داده شده است.

افعال مرکب با بودن و شدن بعضاً در یک مدخل آمده‌اند؛ مثلاً باز شدن و افعال مترادف آن تحت مدخل باز بودن یافت می‌شوند. [۰]

فصل سوم

۳-ارائه روش پیشنهادی

۳-۱ معرفی منطق فازی

منطق فازی یک مدل منطقی ریاضی است که در آن حقیقت می‌تواند نسبی باشد یعنی می‌تواند مقادیری بین ۰ و ۱ داشته باشد. این منطق بر اساس استدلال تقریبی به‌جای استدلال دقیق است.

۳-۱-۱ مجموعه‌های فازی

یک مجموعه فازی، مجموعه‌ای است که اجازه می‌دهد اعضای آن، درجه عضویت متفاوتی در بازه [۰،۱] داشته باشند. در منطق کلاسیک، عضویت یک عضو از یک مجموعه با صفر (۰) نمایش داده می‌شود اگر به مجموعه تعلق نداشته باشد؛ با یک (۱) نشان داده می‌شود اگر به مجموعه تعلق داشته باشد؛ یعنی به‌صورت مجموعه {۱، ۰} نشان داده می‌شود، ولی در منطق فازی این مجموعه به‌صورت بازه [۰،۱] توسعه داده شده است.

یک مجموعه فازی توسعه‌ای از مجموعه کلاسیک است. اگر X، جهان مورد بحث باشد و اعضای آن با x نشان داده شوند، آنگاه مجموعه فازی A از X با زوج مرتب با رابطه زیر تعریف می‌شود:

$\mu_A(x)$ تابع عضویت، x در A است.

$$A=\{(x, \mu_A(x))|x \in X\}$$

اعداد فازی روشی برای توصیف عدم دقت و ابهام داده هستند. یک عدد فازی در مفهوم توسعه‌ای از یک عدد منظم است که به مقدار منفرد اشاره نمی‌کند، اما تا حدودی با مجموعه مقادیر ممکن ارتباط برقرار می‌کند. این مقدار برای خودش یک وزن بین '۰' و '۱' دارد. این وزن تابع عضویت نامیده می‌شود.[۴۴]

۳-۱-۲ منطق فازی در مقابل احتمال

احتمال می‌تواند به دو دسته وسیع اسنادی[۲۲] و فیزیکی[۲۳] دسته‌بندی شود. رویداد E با احتمال ۸۰٪ در نظر بگیرید احتمال اسنادی می‌گوید با مدرک موجود فعلی درست نمایی رخداد E ۸۰٪ می‌باشد. به‌عبارت‌دیگر احتمال فیزیکی می‌گوید که اگر آزمایش شامل E به تعداد دفعات نامتناهی انجام شود آنگاه در ۸۰٪ موارد E اتفاق خواهد افتاد.

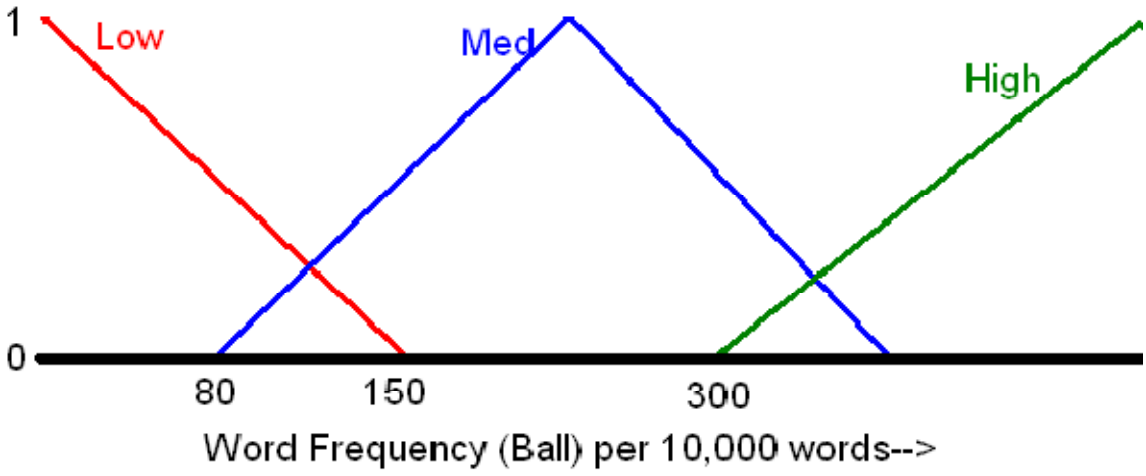
برای مثال خودرو A که ۱۵۰ کیلومتر در ساعت حرکت می‌کند. حال این خودرو سریع تعریف می‌شود. در نظر بگیرید خودرو B با سرعت ۱۸۰ کیلومتر در ساعت حرکت می‌کند. در یک آزمایش تکرارشونده چنین نیست که خودرو B سریع‌تر از A باشد (احتمال فیزیکی). محتمل‌تر است که خودرو B سریع‌تر از خودرو A باشد (احتمال اسنادی). در واقع هر دو خودرو ویژگی سریع را دارند اما با درجات مختلف.

اگر برای این مسئله منطق فازی را استفاده کنیم و کلاسی به نام خودروهای سریع وجود داشته باشد که شامل خودروهایی با ویژگی سریع باشد آنگاه مقادیر عضویت هر دو خودرو A و B می‌توانند مقایسه شوند. مقدار عضویت بالاتر بدین معنی است که خودرو مربوطه درجه بالاتری از ویژگی سریع را داراست.[5]

۳-۱-۳ درجه درستی

درجات درستی به وسیله تابع عضویت تعیین می‌شوند مثال در شکل زیر نشان داده شده است. [5]

[22] evidential
[23] Physical



شکل ۳-۱ توابع عضویت فازی برای تکرار واژه «ball»

با توجه به شکل بالا فرض کنید به ما گفته شده که متنی که در آن تکرار لغت «ball» در آن زیاد باشد مربوط به ورزش است. به منظور نسبت دادن مقادیر عددی به تکرار لغات «کم»، «متوسط» و «زیاد» از تابع عضویت استفاده می‌شود.[5]

برای مثال اگر تعداد تکرار لغت "ball" ۳۰۰ باشد آنگاه آن متن با درجه عضویت حدود ۰.۵ عضو مجموعه Med است و با درجه عضویت صفر عضو مجموعه high است؛ به عبارت دیگر با تعداد تکرار ۳۰۰ زیاد نیست ولی تا حدودی متوسط است و به هیچ وجه کم نیست زیرا درجه عضویت برای low صفر است.

۳-۲ مراحل متن‌کاوی توسط منطق فازی

مراحل ذیل جهت به‌کارگیری منطق فازی در متن‌کاوی باهدف تفکیک و دسته‌بندی موضوع انجام می‌شود.

۳-۲-۱ پیش پردازش متن

این فاز شامل تصفیه متن مثل حذف تبلیغات و غیره می‌باشد. همچنین مجبور به پاک سازی تگ‌ها و خط تیره‌ها از متونی مثل HTML و XML هستیم. حذف این متن ناخواسته به کارایی الگوریتم کمک می‌کند.

۳-۲-۲ تولید مشخصه[۲۴]

اسناد به روش کیسه لغات نمایش داده می‌شوند. حرف‌اضافه‌ها حذف می‌شوند. حذف پیشوند و پسوند لغات انجام می‌شود. این کار باعث نمایش ریشه لغت می‌شود. برای مثال افعال با زمان‌های مختلف به یک زمان تبدیل می‌شوند تا از این جهت تفکیک بین متون انجام نشود.

۳-۲-۳ انتخاب ویژگی[۲۵]

در اینجا مشخصه‌هایی که استفاده خواهد شد انتخاب می‌شوند. این انتخاب مشخصه‌ها چه قبل و چه هنگام استفاده می‌توانند انجام شوند. باید مشخصه‌هایی انتخاب شوند که کارایی لازم را داشته باشند برای مثال به‌منظور دسته‌بندی متون ورزشی و سیاسی در نظر گرفتن وجود اسامی اشخاص در متن به‌عنوان مشخصه کمکی نخواهد کرد زیرا هر دو متن از این جهت مشترک هستند در عین حال وجود لغاتی مانند استادیوم، توپ و غیره می‌تواند نشان دهنده متن‌هایی در زمینه ورزش باشد و به‌طور مشابه لغاتی مانند انتخابات، دموکراسی و مجلس مرتبط با متن‌های با موضوع سیاسی است. با این وجود در بعضی از موقعیت‌ها اشاره به مشخصه‌هایی که بین متون تمییز دهد مشکل است. در چنین موقعیت‌هایی نیاز به یافتن مشخصه‌هایی است که به دسته‌بندی کمک کنند بدین منظور الگو برداری از متون که در دسته‌های مشخص دسته‌بندی شده و آموزش آن‌ها به سیستم هوشمند برای یادگیری و تفکیک متون هنگام تست کمک می‌کند. تکرار لغات برای لغات مختلف برای چنین متن‌هایی می‌تواند مشخصه‌ای جهت جداسازی متون از یکدیگر باشد.

---

[24] Feauture generation
[25] Fearture selection

۳-۲-٤ دسته‌بندی[۲۶]

در روش دسته‌بندی FCM[۲۷] هر متنی که باید دسته‌بندی شود به صورت کیسه لغات نمایش داده می‌شود و از آن کیسه تنها لغات ضروری (مشخصه‌ها) نگه داشته می‌شوند. اکنون هدف دسته‌بندی متن‌ها به بخش‌های مختلف (کلاس‌ها در FCM) است. با استفاده از FCM متن‌هایی که دارای تکرار لغات مشابه (با توجه به طول متن) از مشخصه‌های منتخب هستند در یک دسته یا گروه قرار می‌گیرند.

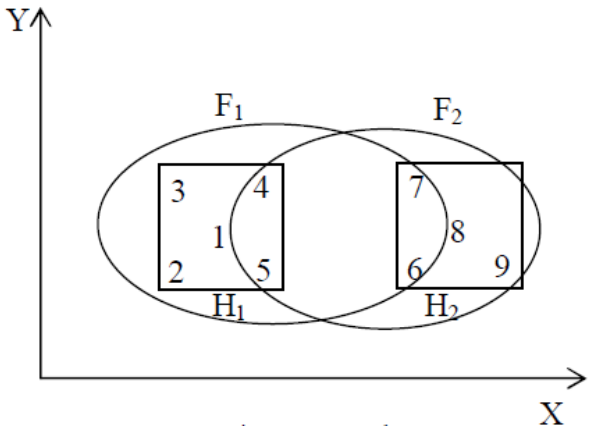
۳-۲-۵ سنجش و ارزیابی نتایج

به‌وسیله درجه داده شده به یک متن جهت تعلق به یک دسته از متون می‌توان نسبت تعلق متن به آن دسته را مشخص کرد یعنی از دید فازی هرچه درجه عضویت یک متن برای دسته داده شده بالاتر باشد تعلق متن به آن دسته بیشتر است. با این وجود اگر تمام مقادیر عضویت مرتبط با یک متن خاص یکسان باشد یا خیلی مشابه باشد و نتوان به‌خوبی تشخیص داد که متن جزء کدام دسته است چنین تصور می‌شود که متن متعلق به هیچ کدام از این دسته‌ها نیست.

دسته‌بندی متن توسط منطق فازی نه تنها نام دسته بلکه درجه عضویت متن به آن دسته نیز به دست می‌آید. [5]

شکل ۳-۲ تفاوت خوشه بندی فازی و خوشه بندی دقیق را نشان داده است. برای مثال در این شکل الگوی شماره ۴ با درجه عضویت بالایی عضو خوشه F1 و با درجه عضویت کمی عضو خوشه F2 می‌باشد ولی از دید خوشه بندی دقیق و سخت‌گیرانه الگوی ۴ فقط عضو خوشه H1 می‌باشد.

[26] classification
[27] Fuzzy C-means



شکل ۳-۲ خوشه‌های فازی

۳-۳ معرفی مرورگر فارس نت

فارس نت یا همان WordNet فارسی نرم افزار الهام گرفته از WordNet حاوی بانک اطلاعاتی از نوع xml و غنی از لغات است که لغات را در دسته‌های هم معنی بخش‌بندی کرده است.

آمار یافت شده از بانک اطلاعاتی فارس نت به صورت زیر است:

- تعداد synset ها: ۱۰۰۳۰ رکورد
- فایل xml با نام SENSES_RELATIONS_Map1: ۳۶۰ رکورد
- تعداد لغات: ۱۷۰۳۴
- فایل xml با نام SYNSET_RELATIONS_Map: ۶۷۳۸ رکورد

جداول زیر همان فایل‌های Xml هستند که پایگاه داده فارس نت را تشکیل می‌دهند؛ و فارس نت با استفاده از پرسشگری از فایل‌های xml نتایج دلخواه را به کاربر به شکل گرافیکی و کاربرپسند ارائه می‌دهد.[45]

در جدول ۳-۱ قسمتی از فایل SENSES_RELATIONS_Map1 .xml که نمایی از بانک اطلاعاتی ارتباطات بین sense هاست نمایش داده شده است. ستون WID_1[28] و WID_2 شامل لغات هستند

[28] Word identifier

که با شناسه در این جدول نشان داده شده‌اند. هر رکورد از جدول، نوع ارتباط[۲۹] بین WID_1 و WID_2 را مشخص می‌کند. برای مثال رکورد اول از جدول بیان می‌کند که لغات با شناسه ۱۰۶۴۲ و ۱۰۶۸۲ متضاد[۳۰] یکدیگر هستند.

جدول ۳-۱ نمایی از فایل SENSES_RELATIONS_Map1 .xml

STATUS	RELATION_TYPE	SID_2	WID_2	SID_1	WID_1
DEFAULT	Antonym	0	10682	1	10642
DEFAULT	Antonym	0	11144	0	10568
DEFAULT	Antonym	0	12870	0	12869
DEFAULT	Antonym	0	12875	0	12874
DEFAULT	Antonym	0	12877	0	12878
DEFAULT	Antonym	0	9828	0	12879

جدول ۳-۲ فایل xml دیگری از پایگاه داده لغوی فارس نت را نشان می‌دهد که نمایانگر synset ها می‌باشد. این جدول نشان می‌دهد که هر synset شامل چه لغاتی است، برای مثال در جدول ۱-۲ سه لغت با شناسه‌های ۹۸۷۴ و ۱۰۲۱۲ و ۱۰۲۱۳ در یک synset قرار دارند.

جدول ۳-۲ نمایی از Synsets_Map1.xml

STATUS2	SID[33]	WID	Type	OFFSET	POS[32]	UID[31]	ID
				948103	Adjective	1000000037	1
			Equivalence		Adjective	1000000037	1
DEFAULT	0	9874			Adjective	1000000037	1
DEFAULT	0	10212			Adjective	1000000037	1
DEFAULT	0	10213			Adjective	1000000037	1

[29] Relation type
[30] Antonym
[31] Unique identifier
[32] Part of speech
[33] Synset indentifier

				948670	Adjective	1000000038	2
			Equivalence		Adjective	1000000038	2
DEFAULT	0	10379			Adjective	1000000038	2
DEFAULT	0	10429			Adjective	1000000038	2
				1251128	Adjective	1000000039	3

جدول ۳-۳ ادامه و انتهای فایل Synsets_Map1.xml

				1819147	Verb	10017	10017
			Equivalence		Verb	10017	10017
DEFAULT	0	17853			Verb	10017	10017
DEFAULT	1	17854			Verb	10017	10017
					Adjective	10029	10029
DEFAULT	1	17862			Adjective	10029	10029
					Adjective	10030	10030
DEFAULT	2	17862			Adjective	10030	10030
					Adjective	10030	10030
DEFAULT	2	17862			Adjective	10030	10030

جدول ۳-۴ نیز قسمتی از فایل xml شامل لغات را نشان می‌دهد. در ستون اول شناسه لغت و در ستون دوم مقوله لغوی و در ستون‌های سوم و چهارم به ترتیب مقدار یا همان ماهیت لغت و آوای تلفظی لغت آمده است.

جدول ۳-۴ نمایی از فایل Words_Map.xml

STATUS3	POS2	CrID	SID	STATUS	AVAINFO	WORDVALUE	POS	WID
DEFAULT	Noun	2	0	DEFAULT	yekhezAr	1000	Noun	12893
DEFAULT	Noun	2	0	DEFAULT	yekmilyArd	1000000000	Noun	3
DEFAULT	Noun	2	0	DEFAULT	X	۲۴ساعت	Noun	4
DEFAULT	Noun	2	1	DEFAULT	X	۲۴ساعت	Noun	4
DEFAULT	Noun	2	0	DEFAULT	X	۳۶۵روز	Noun	5
DEFAULT	Noun	2	0	DEFAULT	X	۶۰دقیقه	Noun	6
DEFAULT	Noun	2	0	DEFAULT	'AbAn	آبان	Noun	7

DEFAULT	Noun	2	0	DEFAULT	'AbAnmAh	آبان ماه	Noun	8
DEFAULT	Noun	2	0	DEFAULT	vasileye azAdAri	وسیله عزاداری	Noun	9
DEFAULT	Noun	2	0	DEFAULT	'AmuzeSi	آموزشی	Noun	10
DEFAULT	Noun	2	0	DEFAULT	'Amper	آمپر	Noun	11
DEFAULT	Noun	2	0	DEFAULT	'AvAzxAn	آوازخوان	Noun	12
DEFAULT	Noun	2	0	DEFAULT	'Aye	آیه	Noun	13

در جدول ۳-۵ که برگرفته از فایل SYNSET_RELATIONS_Map.xml است، ارتباطات بین synset ها آمده است.

جدول ۳-۵ نمایی از فایل SYNSET_RELATIONS_Map.xml

STATUS	RELATION_TYPE	ID_2	ID_1
DEFAULT	Hypernym	5423	2526
DEFAULT	Hypernym	4825	4737
DEFAULT	Hypernym	6052	4737
DEFAULT	Hypernym	4276	2530
DEFAULT	Hypernym	5801	2531
DEFAULT	Hypernym	5218	6445
DEFAULT	Hypernym	6445	6450
DEFAULT	Hypernym	6450	2532
DEFAULT	Hypernym	3285	6450
DEFAULT	Hypernym	5380	6450
DEFAULT	Hypernym	4472	6450

برای مثال در رکورد اول از جدول ۳-۵ ارتباط بین synset ها با شناسه ۲۵۲۶ و ۵۴۲۳ از نوع Hypernym است.

در جدول بالا ستون RELATION_TYPE شامل یکی از موارد زیر است به‌عبارت‌دیگر ارتباط بین لغات در موارد زیر خلاصه می‌شود:

1. Hypernym
2. Synonym
3. Antonym
4. Holonym(part of)

5. Causes
6. Holonym(portion of)
7. Holonym(member of)
8. Has Mero made of
9. Hyponym

معانی و تعریف واژگان بالا (انواع ارتباط بین synset ها) در پیوست آمده است.

مراحل بارگذاری فایل‌های پایگاه داده xml برای مرورگر فارس نت در شکل‌های زیر مشهود است.

در شکل ۳-۳ بارگذاری synset ها که با شناسه و مقوله نحوی مشخص شده، نشان داده می‌شود. همان‌طور که در شکل مشهود است هر synset خود شامل یک یا چند sense می‌باشد.

```
SYNSET (2663 , Noun)
        SENSE[0] : 1190 , 0
        SENSE[1] : 2062 , 0
SYNSET (2664 , Noun)
        SENSE[0] : 4795 , 0
SYNSET (2665 , Noun)
        SENSE[0] : 2055 , 1
        SENSE[1] : 3173 , 0
        SENSE[2] : 3981 , 1
        SENSE[3] : 4050 , 5
        SENSE[4] : 4089 , 0
        SENSE[5] : 4795 , 2
```

شکل ۳-۳ بارگذاری دسته‌های معنایی در پایگاه داده فارس نت.

در جدول زیر قسمتی از فایل بانک اطلاعاتی xml مرتبط با شکل ۳-۳ نشان داده شده است.

جدول ۳-۶ بارگذاری پایگاه داده فارس نت مرتبط با شکل ۳-۳

Synsets_Map1							
STATUS2	**SID**	**WID**	**Type**	**OFFSET**	**POS**	**UID**	**ID**
					Noun	5100000346	2665
DEFAULT	1	2055			Noun	5100000346	2665
DEFAULT	0	3173			Noun	5100000346	2665
DEFAULT	1	3981			Noun	5100000346	2665
DEFAULT	5	4050			Noun	5100000346	2665
DEFAULT	0	4089			Noun	5100000346	2665
DEFAULT	2	4795			Noun	5100000346	2665

```
initSensesRelation() >> 10047 , 0 , 13134 , 0
initSensesRelation() >> 9496 , 0 , 13225 , 0
initSensesRelation() >> 9496 , 0 , 13145 , 0
initSensesRelation() >> 819 , 0 , 13191 , 0
initSensesRelation() >> 10910 , 0 , 10611 , 0
initSensesRelation() >> 10270 , 0 , 11094 , 0
initSensesRelation() >> 10270 , 0 , 13162 , 0
initSensesRelation() >> 9808 , 0 , 11140 , 0
initSensesRelation() >> 9343 , 0 , 13169 , 0
initSensesRelation() >> 9343 , 0 , 12422 , 0
initSensesRelation() >> 10090 , 0 , 12877 , 0
initSensesRelation() >> 9879 , 0 , 10715 , 0
initSensesRelation() >> 13224 , 0 , 13227 , 0
initSensesRelation() >> 10526 , 0 , 13227 , 0
initSensesRelation() >> 3591 , 0 , 9346 , 0
initSensesRelation() >> 12053 , 0 , 13066 , 0
initSensesRelation() >> 10488 , 0 , 7574 , 0
initSensesRelation() >> 10262 , 0 , 9664 , 0
initSensesRelation() >> 17587 , 0 , 10518 , 1
initSensesRelation() >> 17588 , 0 , 10518 , 1
initSensesRelation() >> 17592 , 0 , 17590 , 0
initSensesRelation() >> 17606 , 0 , 17605 , 0
initSensesRelation() >> 11125 , 0 , 10852 , 0
initSensesRelation() >> 11125 , 0 , 10852 , 0
```

شکل۳-۴- بارگذاری sense relation ها

جدول ۳-۷ پایگاه داده متصل به شکل۳-۴ ارتباط متضاد واژگان با شناسه‌های ۱۷۵۹۰ و ۱۷۵۹۲ مشهود است.

SENSES_RELATIONS_Map1					
STATUS	RELATION_TYPE	SID_2	WID_2	SID_1	WID_1
DEFAULT	Antonym	0	17590	0	17592
DEFAULT	Antonym	0	17605	0	17606

مثال:

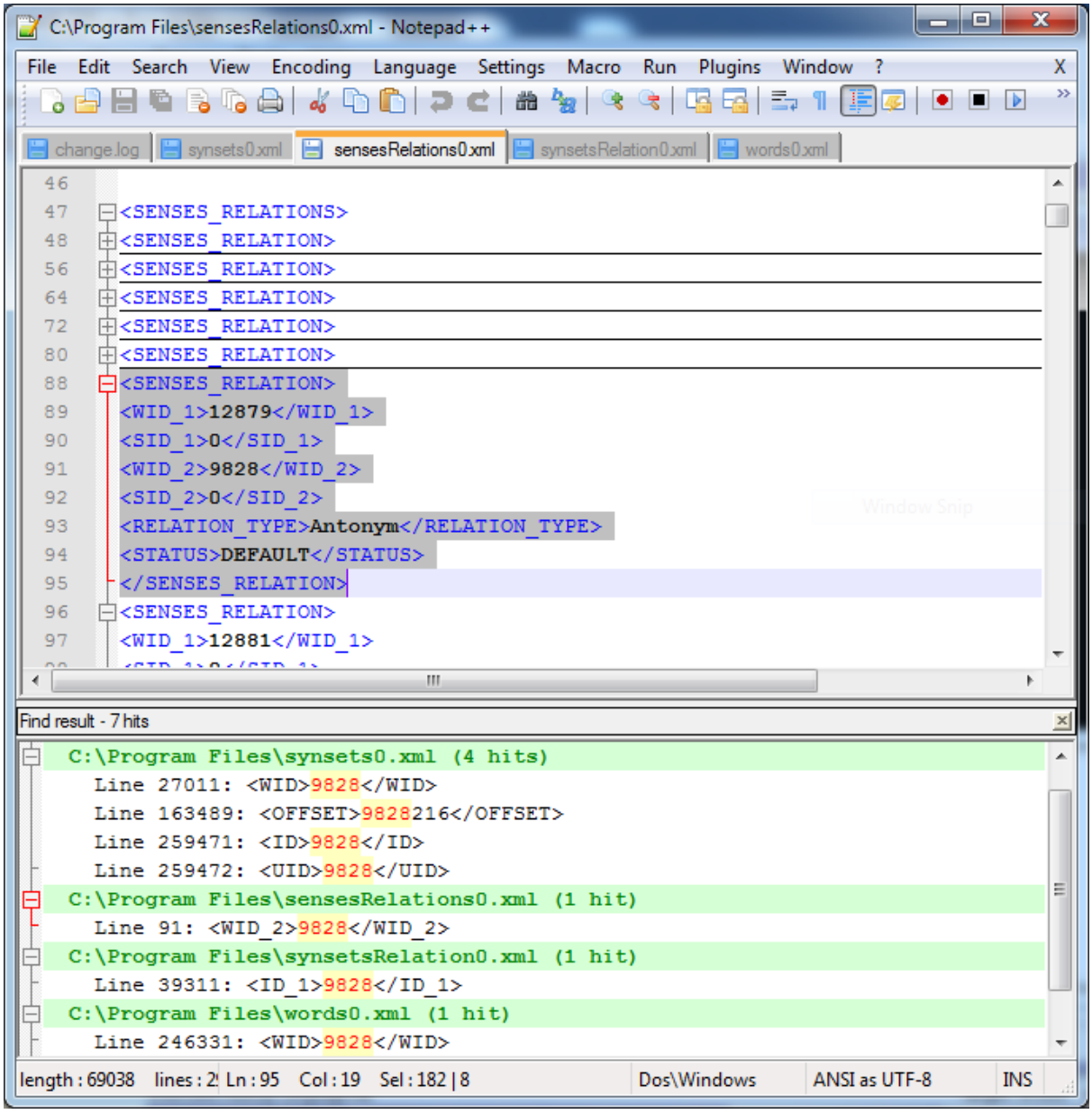
برای مثال دو واژه «باحوصله» و «بی‌حوصله» در پایگاه داده فارس نت در نظر گرفته می‌شود که به صورت شکل‌ها و جداول زیر نمود دارد:

در شکل ۳-۵ نشان داده شده که این دو لغت هر دو صفت هستند.

شکل ۳-۵ Words_Map.xml دارای دو واژه «باحوصله» و «بی‌حوصله»

AVAINFO	WORDVALUE	POS	WID
bi hosele	بی حوصله	Adjective	**9828**
bahosele	باحوصله	Adjective	**12879**

در شکل ۳-۶ نحوه نمایش متفاوت فایل xml از ارتباطات sense های دو واژه «باحوصله» و «بی‌حوصله» است.



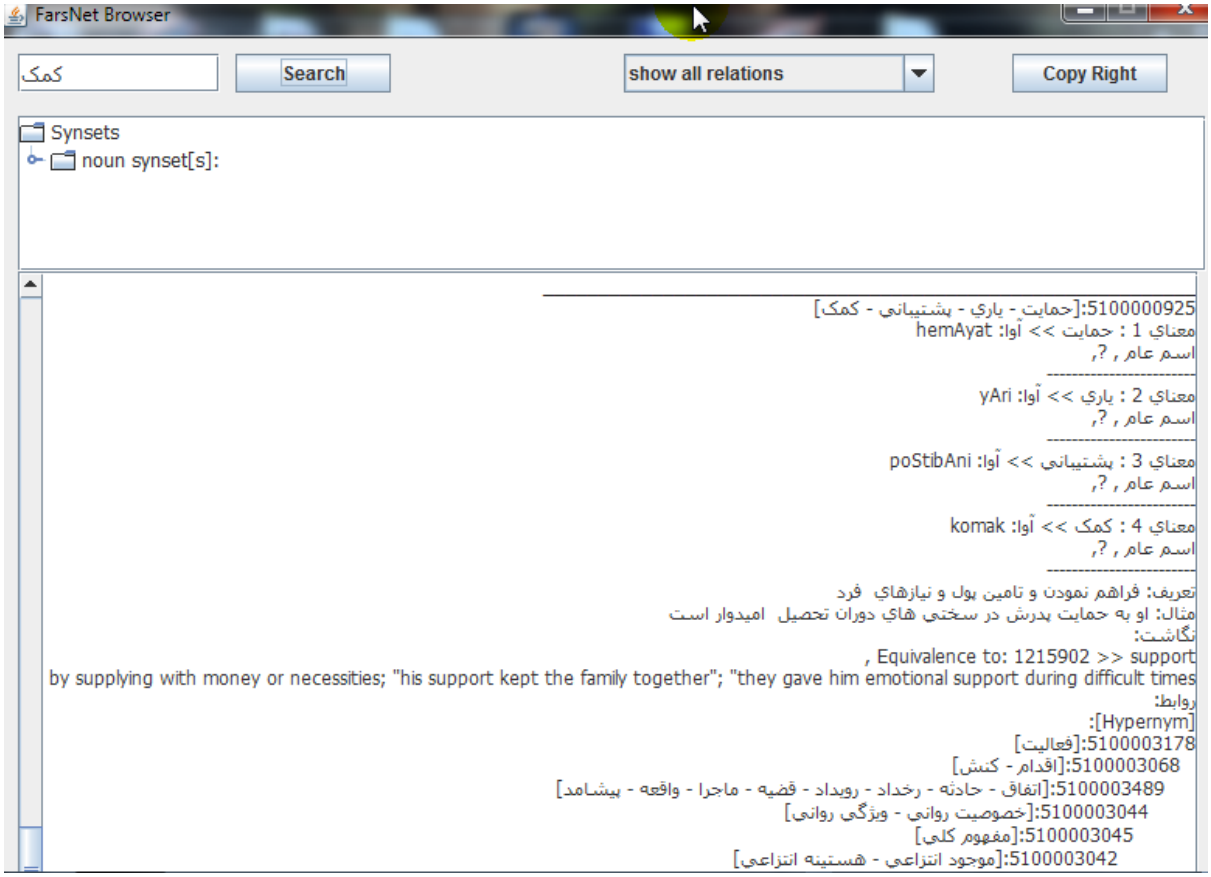
شکل ۳-۶ دو واژه «باحوصله» و «بی‌حوصله» در قالب شناسه‌های خود مرتبط هستند.

در جدول ۳-۸ فایل xml شکل ۳-۶ برای دو واژه نشان داده شده است.

جدول ۳-۸ SENSES_RELATIONS_Map1.xml نشان دهنده دو واژه «باحوصله» و «بی‌حوصله»

STATUS	RELATION_TYPE	SID_2	WID_2	SID_1	WID_1
DEFAULT	Antonym	**0**	**9828**	**0**	**12879**

نمونه‌ای از عملکرد این نرم افزار در شکل۳-۷ مشاهده می‌شود.



شکل۳-۷- جستجوی واژه «کمک» در FarsNet Browser

در شکل۳-۷ برای واژه «کمک» synset های مربوط به این واژه آورده شده است و بعد از آن تعریف، مثال و معادل این واژه در پایگاه داده لغوی WordNet ذکر شده است. در انتها رابطه از نوع Hypernym با دیگر synset ها که با شناسه مشخص شده آمده است.

۳-۴ فازنت

۳-٤-۱ پایگاه داده لغوی فازنت

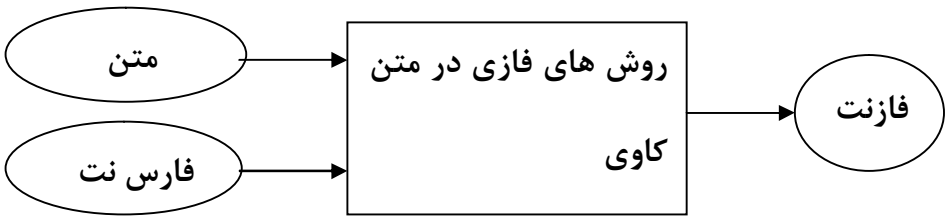
هدف، استفاده از ارتباط فازی برای اضافه کردن به ارتباطات در پایگاه داده لغوی فارس نت است. اینکه این ارتباط چطور عدد بگیرد نیاز به معیار نمره دهی است. با توجه به کاربرد فازی در این زمینه از امکان سنجش آماری استفاده می‌شود.

۳-٤-۲ تعریف فازنت

برنامه پیشنهادی پایگاه داده لغوی FuzzNet است. توصیف فرمولی این پایگاه داده به صورت زیر است:

FuzzNet=FarsNet+Fuzzy=Dictionary+sense+Relation+Fuzzy

هدف پایان نامه در قالب فرمول فازنت=فارس نت + فازی خلاصه می‌شود که در شکل زیر به عنوان خروجی فرآیندی است که با استفاده از روش‌های فازی در متن‌کاوی و ورودی متن + فارس نت، فازنت را تولید می‌کند.



شکل ۳-۸ روند کلی تولید فازنت

۳-٤-۳ تفاوت فازنت با فارس نت

در واقع فارس نت به ازای یک واژه انواع معناهایی که آن واژه می‌تواند داشته باشد ارائه می‌دهد اما اینکه کدام معنا برای واژه بیشتر به‌کارگیری می‌شود یا دارای اولویت بالاتر است سکوت می‌کند و در عمل تفاوتی بین معناها یا همان sense ها قائل نیست.

برای اینکه بتوان اولویت بندی sense ها را برای هر واژه اجرا کرد باید معیاری برای نمره دهی و امتیازبندی لغت وجود داشته باشد. روش پیشنهادی برای تعیین اولویت هر sense برای یک واژه آمارگیری تعداد کاربرد آن sense در متن‌های مختلف است یعنی برای مثال هرچه نسبت تعداد یک sense به دیگر sense ها برای واژه مورد بررسی بیشتر باشد اولویت آن sense بالاتر است. با وجود آموزش از متون مختلف، این نمره‌دهی با استفاده از آموزش یک سیستم هوشمند به واقعیت نزدیک‌تر خواهد بود و بدیهی است با داشتن متن‌های بیشتر نتیجه اولویت بندی به واقعیت نزدیک‌تر خواهد بود.

۳-٤-٤ آموزش سیستم فازنت

برای آموزش فازنت به ازای هر واژه متنی غنی از لغات در نظر گرفته می‌شود. به ازای هر بار به‌کارگیری این واژه در متن، sense آن واژه تعیین شده و مورد محاسبه قرار می‌گیرد. در انتها نسبت هر sense به تمام sense ها محاسبه می‌شود که این نسبت عددی بین ۰ و ۱ است و به عنوان درجه عضویت فازی آن sense به مجموعه فازی واژه در نظر گرفته می‌شود؛ به عبارت دیگر به ازای هر واژه یک مجموعه فازی تشکیل می‌شود که شامل sense های آن واژه و درجه عضویت فازی هرکدام از sense ها می‌باشد.

مثال:

مثال واضح برای آموزش فازنت آموزش واژه «شیر» است.

در ابتدا sense ها یا همان دسته‌های معنایی برای این واژه از پایگاه داده فارس نت استخراج می‌شود.

دسته معنایی شماره یک (sense1): شیر خوردن

دسته معنایی شماره دو (sense2) : شیر آب

دسته معنایی شماره یک (sense3) : شیر جنگل

واژه شیر را در پایگاه داده لغوی آموزش جستجو می‌کنیم به نتایج زیر می‌رسیم.

جدول ۳-۹ نسبت دسته‌های معنایی مختلف برای واژه «شیر» بدون تفکیک متن

انواع sense برای واژه «شیر»	تعداد تکرار هر sense	نسبت تکرار هر sense
Sense1	۵۰	۵۰/۶۲
Sense2	۱۰	۱۰/۶۲
Sense3	۲	۲/۶۲
مجموع	۶۲	۱

با توجه به نتایج جدول بالا مجموعه فازی به صورت زیر تشکیل می‌شود بطوریکه نسبت همان درجه عضویت در نظر گرفته می‌شود.

مجموعه فازی واژه «شیر» = {(Sense1, 0.80), (sense2, 0.16), (sense3, 0.04)}

بدین وسیله با جستجوی واژه «شیر» در پایگاه داده لغوی پیشنهادی نتایج جدول ۳-۱۰ حاصل می‌شود:

جدول ۳-۱۰ نتایج جستجوی واژه «شیر» در پایگاه داده لغوی پیشنهادی

دسته معنایی شماره یک (sense1): شیر خوردن درجه عضویت: ۰.۸۰

دسته معنایی شماره دو (sense2) : شیر آب درجه عضویت: ۰.۱۶
دسته معنایی شماره یک (sense3) : شیر جنگل درجه عضویت: ۰.۰۴

بدین وسیله فازنت به ازای واژه شیر تنها به ارائه دسته‌های معنایی آن اکتفا نمی‌کند و درجه عضویت فازی آن معنا به واژه را نیز ارائه می‌دهد. در نتیجه در فایل xml پایگاه داده فارس نت یک فیلد با درجه فازی اضافه می‌شود و فازنت را می‌سازد.

با در نظر گرفتن دسته بندی موضوعی متون و آمارگیری لغت برای هر دسته بندی موضوعی و محاسبه نسبت‌های تکرار واژگان، به ازای هر موضوع برای هر واژه مجموعه فازی متفاوتی تشکیل داد که هر مجموعه فازی متفاوت موجب درجه فازی متفاوت است و این درجه فازی با دید تخصصی و موضوعی به واژه نگاه می‌کند و در قالب یک فیلد به پایگاه داده XML فارس نت اضافه شده و فازنت را می‌سازد.

فصل چهارم

۴- معرفی نرم افزار

۴-۱ ساختار پایگاه داده WordNet

حال با نگاه جزئی‌تر به بانک اطلاعاتی موجود در نرم افزار WordNet می‌پردازیم.

دو فایل پایگاه داده برای WordNet به قرار زیر است:

۱- wn_LemmaIndexes.lst

۲- wn_database.lst

این دو فایل بوسیله نرم افزار Visual Studio.Net یا یک ویرایشگر متنی قابل مشاهده است.

در فایل wn_LemmaIndexes.lst لغات با ترتیب دیکشنری در آن جمع آوری شده است و برای هر لغت یک یا چند عدد[۳۴] حداکثر ۷ رقمی آورده شده است که شماره شناسایی synset محسوب می‌شود.

لغاتی که در یک synset مشترک باشند عدد جلوی آن‌ها یکسان است.

مثال:

hello 6269324
hi 626932,4507692
how-do-you-do 6466191,6269324
howdy 6269324
hullo 6269324

به ازای هر واژه، شناسه synset هایی که آن واژه به آن‌ها تعلق دارد نوشته شده است که می‌تواند یک یا چند synset باشد.

همان‌طور که در بالا آمده است عدد 6269324 بین لغات بالا مشترک است و به‌عبارت‌دیگر لغات بالا در یک synset قرار دارند و قرابت معنایی با هم دارند.

در فایل wn_database.lst برخلاف فایل پیش گفته بر اساس synset ها مرتب‌سازی و دسته‌بندی شده است.

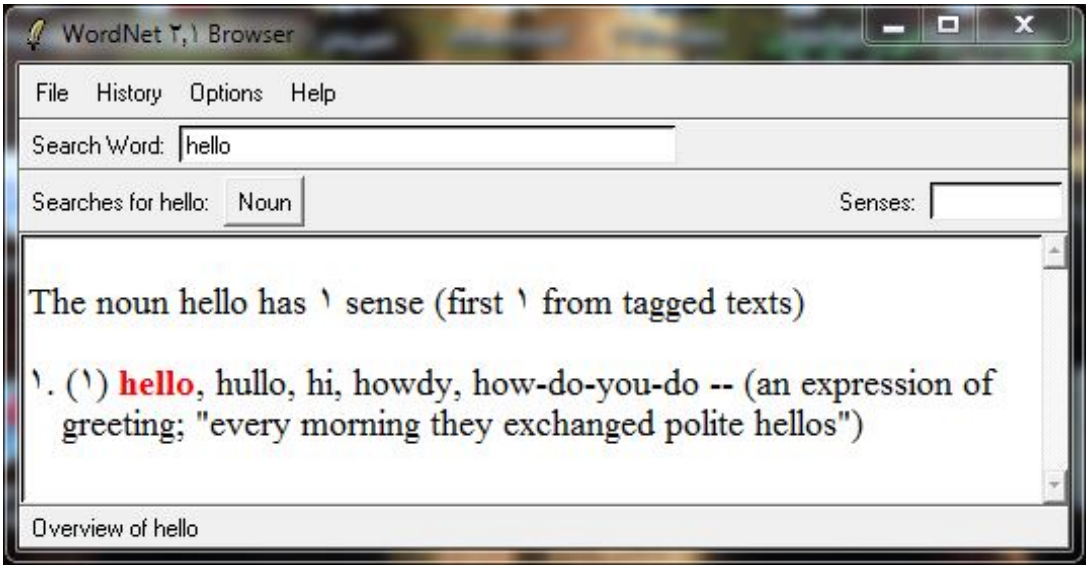
[۳۴] بدیهی است که یک لغت می‌تواند به دو یا چند synset تعلق داشته باشد.

برای مثال برای synset شامل لغت hello در این فایل متن زیر آمده است:

hello,hullo,hi,howdy,how-do-you-do|n|an expression of greeting|"every morning they exchanged polite hellos"

لغات مشترک در یک synset با علامت (ویرگول) «,» از یکدیگر تفکیک شده‌اند. |n| نشان دهنده مقوله noun یا اسم بودن لغات synset است. عبارت بعدی توصیف کننده synset و عبارت داخل گیومه "" مثالی از یکی از لغات synset است.

خروجی معادل در برنامه wordnet 2.1 به شکل زیر است:



شکل۴-۱ نمایش WordNet Browser برای واژه hello

۴-۲ فارس نت

نخستین نسخه از فارس نت شامل وردنت فارسی است که توسط آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی و با حمایت مرکز تحقیقات مخابرات ایران ساخته شده است. این محصول در بسیاری کاربردهای پردازش زبان فارسی ازجمله ترجمه ماشینی، خلاصه‌سازی اخبار، جستجو و بازیابی اطلاعات، کشف هرزنامه‌ها، تحلیل اطلاعات متون و رمزنگاری معنایی نقش کلیدی بازی

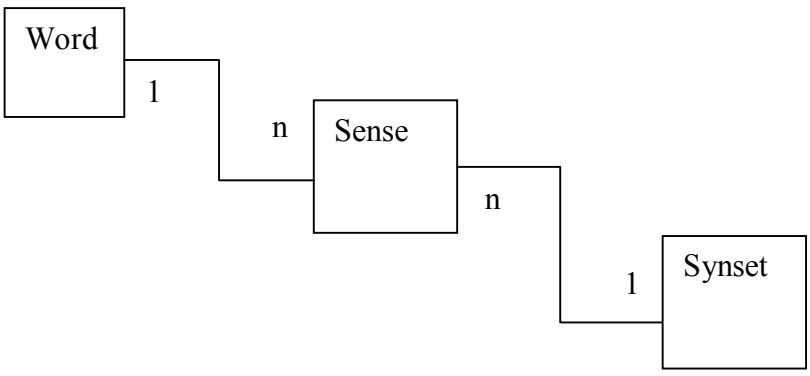
می‌کند. فارس نت دربردارنده زیرمجموعه‌ای از واژگان مورد استفاده در زبان فارسی نوشتاری معیار است که علاوه بر امکان استفاده در سیستم‌های پردازش زبان فارسی امکانات تبدیل دوزبانه را نیز فراهم می‌کند.

٤-٢-١ نسخه‌های مختلف فارس نت

نخستین نسخه از این واژ-هستان شناسی شامل ۱۰،۰۰۰ مجموعه هم معنا و ۱۸،۰۰۰ کلمه فارسی است. کلمات تحت پوشش این محصول دارای ۳ نوع مقوله نحوی (اسم، فعل و صفت) هستند و از بین پررخدادترین کلمات زبان فارسی انتخاب شده‌اند.

فارس نت یک آنتولوژی واژهای در زبان فارسی است این آنتولوژی در آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی به رهبری دکتر مهرنوش شمس فرد ساخته شده است.

نسخه فعلی این آنتولوژی اولین نسخهٔ این محصول بوده و مشتمل بر ۱۰۰۰۰ مجموعه هم معنی (ساین‌ست) و ۱۸۰۰۰ کلمه فارسی است. کلمات تحت پوشش فارس نت در سه نوع نحوی اسم، فعل و صفت موجود هستند. هر کلمه دارای دسته معنایی (ساین‌ست) است که هر یک از آنان شامل مجموعه‌ای از کلمات مترادف (سنس) هستند، بطوریکه یک معنی از کلمه را بیان می‌کنند.[۱]



شکل ۴-۲ ساختار فارس نت

در شکل ۴-۲ ارتباط یک به چند بین لغت و sense و همچنین بین synset و sense در فارس نت نشان داده شده است.[۱]

نسخه دوم فارس نت (وردنت عمومی زبان فارسی) شامل بیش از ۳۰ هزار مدخل واژگانی از مقوله‌های اسم، فعل، صفت و قید می‌باشد. علاوه بر روابط درون-مقوله‌ای مطرح در وردنت انگلیسی (نسخه ۲.۱)، پنج رابطه میان-مقوله‌ای نیز مفاهیم را به هم پیوند می‌دهد و علاوه بر ویژگی‌های در نظر گرفته شده برای واژه‌ها، ویژگی‌های نحوی، ساخت‌واژی و آوایی به واژه‌ها و قاب و ساختار آرگومانی به افعال افزوده شده است. همچنین این وردنت قابلیت اتصال به وردنت‌های دیگر را از طریق نگاشت به وردنت پرینستون نسخه 3,0 داراست. مجموعه فارس نت توسط آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی و با حمایت مرکز تحقیقات مخابرات ایران تهیه شده است. [46]و[47]

فصل پنجم

۵- نتایج آزمایشگاهی

۵-۱ استفاده از فازنت برای ترجمه ماشینی (تست)

در پایگاه داده فارس نت هر واژه عضو یک یا چند دسته معنایی است و به ازای هر دسته معنایی واژه انگلیسی معادل آن معنا به همراه شناسه‌اش در WordNet وجود دارد. درجه عضویت فازی برای هر معنی به ترجمه ماشینی صحیح‌تر کمک می‌کند بدین صورت که سیستم مترجم ماشینی به ازای واژه ورودی مجموعه فازی آن واژه را بررسی می‌کند و دسته معنایی با درجه عضویت فازی بالاتر را انتخاب کرده، کد و واژه معادل آن در WordNet ترجمه لغت ورودی را ارائه می‌دهد.

برای مثال در مجموعه فازی واژه «شیر» به دلیل بالاتر بودن درجه عضویت فازی sense1 این دسته معنایی انتخاب می‌شود که با توجه به درجه عضویت این دسته معنایی دارای صحت حدود ۸۰٪ می‌باشد و بدیهی است که دارای خطای حدود ۲۰٪ است.

حال اگر پایگاه داده لغوی بر حسب موضوع تفکیک شود خطای ترجمه کاهش می‌یابد،

سیستم‌های دسته‌بندی خودکار اسناد با افزایش دسترسی به اسناد الکترونیکی و رشد سریع فضای وب، به رویکردی کلیدی به‌منظور مدیریت اطلاعات و دانش تبدیل گشته‌اند. در دهه‌های اخیر، روش‌های بسیاری به‌منظور دسته -بندی متون پیشنهاد شده است که اکثر آنان بر اساس مدل کیسهٔ لغات هستند

۵-۲ جداسازی و دسته بندی متن

جداسازی و دسته‌بندی متون از لحاظ موضوع در مقاله [5] آمده است. در این مقاله با استفاده از تکنیک‌های فازی و درجه بندی فازی متون مختلف با در نظر گرفتن لغات پر کاربرد هر موضوع جداسازی و دسته‌بندی متون از یکدیگر انجام می‌شود.

توسط منطق فازی علاوه بر دسته‌بندی درجه تعلق به یک دسته خاص نیز مشخص می‌شود یعنی اینکه متن ما با چه درجه‌ای متعلق به دسته[۳۵] یک و با چه درجه‌ای متعلق به دسته دو می‌باشد. دسته‌بندی بر اساس مشخصه‌های[۳۶] متن است.

اگر واژه «شیر» را در متون با موضوعات مختلف جستجو کنیم آمارهای متفاوتی برای جدول Sense های واژه یافت می‌شود. برای مثال اگر واژه شیر را در متن با موضوع جانورشناسی جستجو کنیم انتظار می‌رود به صورت زیر باشد

جدول ۵-۱ نسبت دسته‌های معنایی مختلف برای واژه «شیر» برای متن با موضوع جانورشناسی

انواع sense برای واژه «شیر»	تعداد تکرار هر sense	نسبت تکرار هر sense
Sense1(خوردنی)	۷	۰.۱۶
Sense2(آب)	۰	۰
Sense3(جنگل)	۳۵	۰.۸۴
مجموع	۴۲	۱

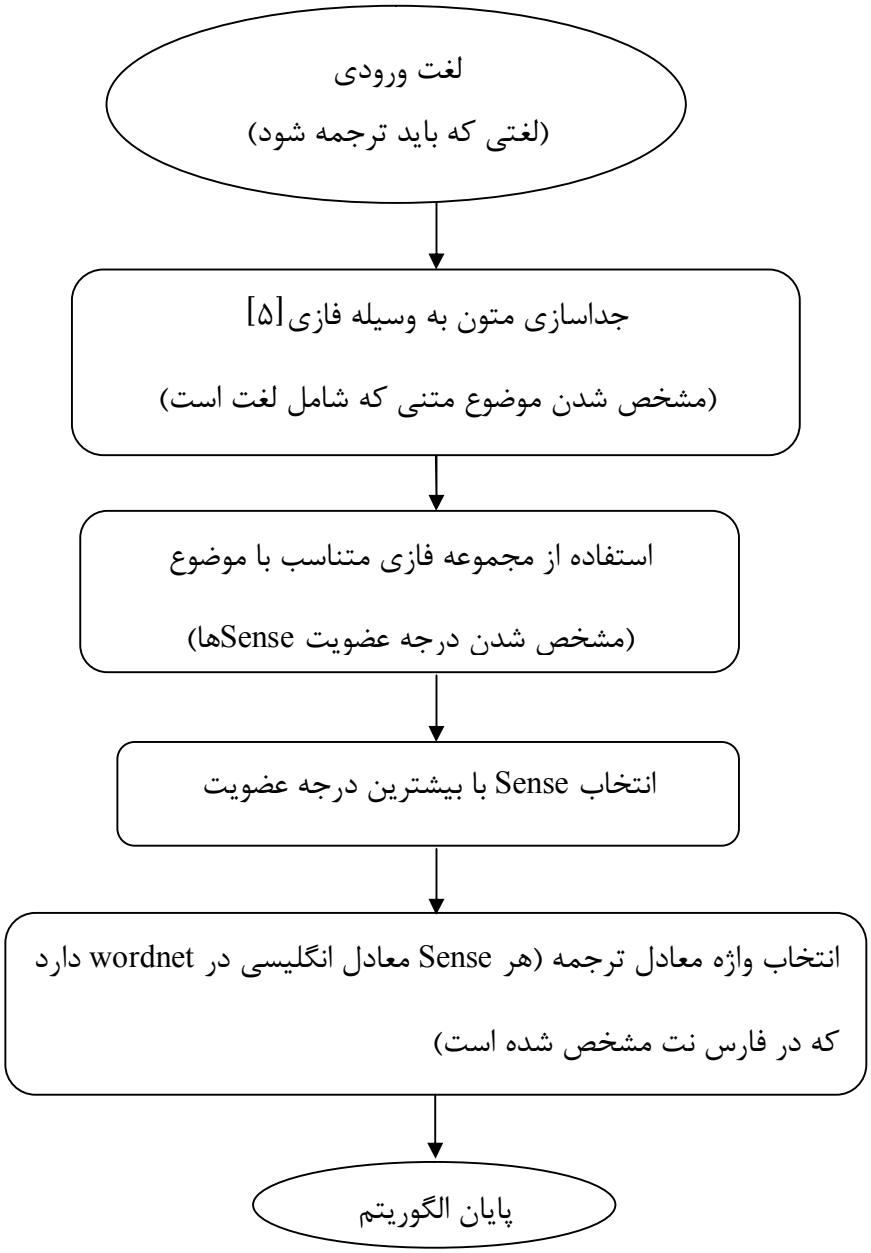
در نتیجه مجموعه فازی برای شیر با درجات عضویت متفاوت به صورت زیر به دست می‌آید

مجموعه فازی واژه «شیر» ={(Sense1, 0.16), (sense2, 0), (sense3, 0.84)}

۵-۳ الگوریتم پیشنهادی

فلوچارت الگوریتم ترجمه لغت با فازنت به صورت زیر است:

[35] Cluster
[36] feature



مثال:

برای الگوریتم ترجمه لغت با فازنت مجموعه فازی برای واژه «شیر» و موضوع متن جانورشناسی به صورت زیر است.

مجموعه فازی واژه «شیر» = {(Sense1, 0.16), (sense2, 0), (sense3, 0.84)}

۵-۴ اجرای الگوریتم

۱- «شیر» ورودی الگوریتم

۲- متن با موضوع زیست شناسی است.

۳- با توجه به موضوع و درجه عضویت Sense های مختلف برای واژه شیر، بالاترین درجه عضویت انتخاب می‌شود.

۴- لذا Sense شماره ۳ انتخاب می‌شود.

۵- sense شماره ۳ با شناسه ۵۱۰۰۰۰۴۵۳۶ در فارس نت ثبت شده و معادل لغت "lion" در WordNet با شناسه ۲۱۲۹۱۶۵ می‌باشد. لذا خروجی الگوریتم واژه "lion" است.

جدول ۵-۲ قسمتی از فایل Synsets_Map1. xml نشان دهنده واژه "lion"

Synsets_Map1							
STATUS2	SID	WID	Type	OFFSET	POS	UID	ID
				2129165	Noun	5100004536	6433
			Equivalence		Noun	5100004536	6433
DEFAULT	0	801			Noun	5100004536	6433
DEFAULT	3	5364			Noun	5100004536	6433

خروجی برنامه مرورگر فارس نت برای Synset بالا به صورت زیر است:

۵۱۰۰۰۰۴۵۳۶:[سلطان جنگل - شیر]

معنای ۱ : سلطان جنگل >>

اسم عام , ?,

معنای ۲ : شیر >> آوا: Sir

اسم عام , ?,

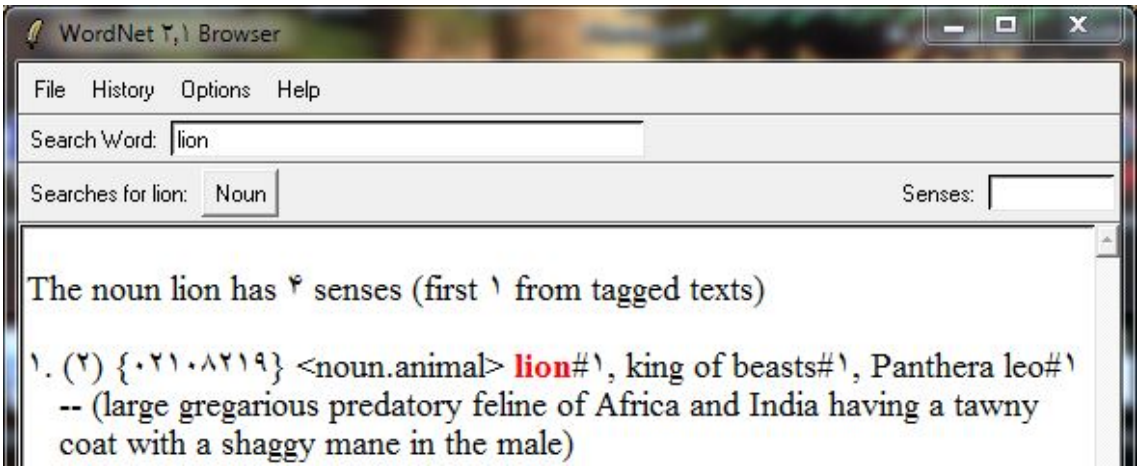
تعریف: جانور پستاندار گوشت خوار بزرگ از گربه سانان، با پشم کوتاه زرد تا خرمایی که جنس نر آن در اطراف سر و گردن یال سیاه یا خرمایی دارد مثال: شیر به طرف گاوها حمله برد نگاشت:
Equivalence to: ۲۱۲۹۱۶۵>> lion ,king_of_beasts ,Panthera_leo, gloss:large gregarious predatory feline of Africa and India having a tawny coat with a shaggy mane in the male
روابط: [Hypernym]: ۵۱۰۰۰۰۰۹۷۱:[حیوان شکارگر - شکارچی - صیاد] ۵۱۰۰۰۰۲۵۳۱:[جانور - حیوان] ۵۱۰۰۰۰۳۱۲۵:[ارگانیسم - سازواره] ۵۱۰۰۰۰۰۵۴۶:[جان دار - جاندار - ذی حیات - موجود زنده] ۵۱۰۰۰۰۳۰۱۹:[شیء - شیئ] ۵۱۰۰۰۰۳۰۴۰:[موجود فیزیکی - هستینه فیزیکی] ۵۱۰۰۰۰۳۰۲۲:[هست - موجود - هستینه]

در عین حال خروجی برنامه WordNet Browserبرای Synset بالا به صورت شکل ۵-۱ است:

۵-۵ ارزیابی الگوریتم فازنت

همان‌طور که در شکل ۵-۱ نشان داده شده است الگوریتم پیشنهادی واژه را به معادل انگلیسی آن ترجمه می‌کند. این الگوریتم با استفاده از ورودی متن به ترجمه بهتر کمک می‌کند، تعیین موضوع

متن توسط دسته‌بندی فازی کمک می‌کند تا الگوریتم برای انتخاب مجموعه فازی هر واژه، متناسب با موضوع اقدام کند.



شکل ۵-۱ خروجی برنامه WordNet Browser برای واژه "lion"

جدول مقایسه ترجمه با الگوریتم فازنت و بدون فازنت برای ترجمه واژه «شیر» در متن با موضوع جانورشناسی به صورت زیر است:

جدول ۵-۳ مقایسه ترجمه با الگوریتم فازنت و بدون فازنت

ترجمه واژه «شیر»	ترجمه ماشینی بدون فازنت		ترجمه ماشینی با فازنت	
	احتمال	واژه	نسبت آماری	واژه
	۰.۳۳	Sense1(خوردنی)	۰.۱۶	Sense1(خوردنی)
	۰.۳۳	Sense2(آب)	۰	Sense2(آب)
	۰.۳۳	Sense3(جنگل)	۰.۸۴	Sense3(جنگل)
درصد خطا	٪۶۶		٪۱۶	
فرمول خطای کلی	$1 - \frac{1}{(\text{تعداد } \textbf{sense} \text{ ها})}$		$1 - \max_i(\text{sense}_i)$	

فصل ششم

۶- نتیجه گیری و پیشنهادات

۶-۱ ارزیابی نتایج

ایجاد مجموعه فازی برای هر واژه و تعیین درجه عضویت هر دسته معنایی به آن واژه با توجه و محاسبه متون عمومی دارای خطای بالاتر نسبت به محاسبه بر اساس متون تخصصی است. بدیهی است در متن عمومی هر واژه می‌تواند بار معنایی گسترده داشته باشد و گرچه با غنای پایگاه داده لغوی برای یافتن مجموعه فازی هر واژه دقت درجات عضویت مجموعه‌های فازی بالاتر می‌رود اما در الگوریتم ترجمه دسته معنایی با بالاترین درجه عضویت انتخاب می‌شود و دسته‌های معنایی با درجات عضویت پایین‌تر نادیده گرفته می‌شود و باعث خطای ترجمه می‌شود.

با وجود تفکیک موضوع متن حاوی واژه‌ای که باید توسط الگوریتم فازنت ترجمه شود دسته‌های معنایی دقیق‌تر خواهد شد زیرا دسته‌های معنی برای متون تخصصی گستردگی کمتری نسبت به متون عمومی دارد و در نتیجه خطای الگوریتم کمتر است.

برای مثال در مجموعه فازی واژه «شیر» به دلیل بالاتر بودن درجه عضویت فازی sense1 این دسته معنایی انتخاب می‌شود که با توجه به درجه عضویت این دسته معنایی دارای صحت حدود ۸۰٪ می‌باشد و بدیهی است که دارای خطای حدود ۲۰٪ است.

حال اگر پایگاه داده لغوی بر حسب موضوع تفکیک شود خطای ترجمه کاهش می‌یابد،

برای مثال اگر برای واژه شیر تفکیک متن نشود در متن عمومی معادل milk انتخاب می‌شود ولی اگر برای واژه شیر موضوع یا همان دسته‌بندی جانورشناسی انجام شود معادل متفاوت "lion" خروجی الگوریتم خواهد بود.

۶-۲ نتیجه گیری و کارهای آینده

برای کارهای آینده می‌توان بر روی اصلاح و کاهش خطای الگوریتم ترجمه فازنت تحقیق کرد برای مثال اگر علاوه بر موضوع متن واژه مورد ترجمه به لغات همسایه آن واژه یا جملات بعد و قبل آن تحلیل و بررسی داشت امید است بتوان با اعمال الگوریتم‌های مناسب به تقویت ترجمه و کاهش خطا پرداخت.

البته ایجاد پایگاه دانش به‌وسیله فارس نت و همزمان اعمال مباحث فازی می‌تواند به ایجاد دائره‌المعارف واژگان فارسی منجر شود و زمینه خوبی برای تحقیق و توسعه است که علاوه بر توسعه علمی به پیشرفت زبان فارسی و همچنین ابزار مفید زبان شناسی نیز تبدیل خواهد شد.

بطور کلی هر چه دید هوشمندانه سیستم مترجم به مفهوم لغت بیشتر و وسیع‌تر باشد بدیهی است خطای کمتری خواهد داشت. این دید وسیع را می‌توان توسط روش‌ها و الگوریتم‌های هوش مصنوعی اعمال کرد که یافتن الگوریتم‌ها و روش‌های مناسب برای متون مختلف نیاز به تحقیق و پژوهش بیشتر دارد. در واقع وقتی چندین روش برای حل یک مسئله وجود دارد پس نتیجه این است که بهترین روش وجود ندارد بلکه متناسب با هر مسئله‌ای باید راه حل مناسب آن را تجویز کرد.

منابع و مآخذ:

[0]- فرهنگ طیفی زبان فارسی- نسخهٔ رقومی [جمشید فراروی. (۱۳۸۷). فرهنگ طیفی. تهران: انتشارات هرمس.]

[1] ص. س. مدنی, ""بهبود دقت سیستم دسته بندی خودکار اسناد فارسی به کمک هستان and ح. حسن پور شناسی فارس نت,"" *پرایانش نرم و فناوری اطلاعات*" vol. 3, pp. 44–54, 1393.

[2] H. L. . I. Roediger, *Artificial Intelligence and Intelligent Systems: The Implications.*, 3rd ed., vol. 35, no. 6. Oxford University Press, 1990.

[3] C. D. Stephens, *Artificial intelligence.*, 3rd ed., vol. 180, no. 8. Tata McGraw-Hill Education Private Limited, 1996.

[4] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *J. Emerg. Technol. Web Intell.*, vol. 1, no. 1, pp. 60–76, Aug. 2009.

[5] S. Goswami and M. S. Shishodia, "a Fuzzy Based Approach To Text Mining and," 2012.

[6] J. Frank, "Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell," *Vasa*, 2006. [Online]. Available: http://medcontent.metapress.com/index/A65RM03P4874243N.pdf\nhttp://link.springer.com/content/pdf/10.1007/978-0-387-69008-7.pdf.

[7] M. V. C. Guelpeli and A. C. B. Garcia, "An analysis of constructed categories for textual classification using fuzzy similarity and agglomerative hierarchical methods," in *Proceedings - International Conference on Signal Image Technologies and Internet Based Systems, SITIS 2007*, Springer, 2007, pp. 92–99.

[8] S. Puri, S. Vector, and M. Classifier, "A Fuzzy Similarity Based Concept Mining Model for Text Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, pp. 115–121, 2011.

[9] H. Hashimi, A. Hafez, and H. Mathkour, "Selection criteria for text mining approaches," *Comput. Human Behav.*, 2015.

[10] R. Ahmad and T. Sim, "Document Topic Generation in Text Mining by Using Cluster Analysis with EROCK," *Int. J. Comput. Sci. Secur.*, vol. 4, no. 4, pp. 176–182, 2008.

[11] E. A. Calvillo, A. Padilla, J. Munoz, J. Ponce, and J. T. Fernandez, “Searching research papers using clustering and text mining,” *23rd Int. Conf. Electron. Commun. Comput. CONIELECOMP 2013*, vol. 4, no. 4, pp. 78–81, 2013.

[12] V. M. Navaneethakumar and C. Chandrasekar, “A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model,” *IJCSI Int. J. Comput. Sci. Issues*, vol. 9, no. 4, pp. 365–370, 2012.

[13] S. K. Rath, M. K. Jena, T. Nayak, and B. Bisoyee, “Data Mining: a Healthy Tool for Your Information Retrieval and Text Mining,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 2, 2011.

[14] P. Senellart and V. D. Blondel, “Automatic Discovery of SimilarWords. Survey of Text Mining II. 2008, 1, 25-44.” Springer London, 2008.

[15] D. Y. K. J. Sumit Vashishta, “Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 4, 2011.

[16] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, I. Müller-Gorman, and A. Zimek, “Finding Hierarchies of Subspace Clusters,” in *Knowledge Discovery in Databases: PKDD 2006*, vol. 4213, Springer, 2006, pp. 446–453.

[17] H.-P. Kriegel, P. Kröger, and A. Zimek, “Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering,” *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 1, pp. 1–58, 2009.

[18] T. Silwattananusarn, “Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 2, no. 5, pp. 13–24, 2012.

[19] A. Krause, J. Leskovec, and C. Guestrin, “Data association for topic intensity tracking,” in *Proceedings of the 23rd international conference on Machine learning - ICML ’06*, 2006, pp. 497–504.

[20] S. Logeswari, “A Survey on Text Mining in Clustering,” *Int. J. Eng. Comput. Sci. ISSN*, no. 0976, pp. 111–116, 2010.

[21] M. F. Caropreso, S. Matwin, and F. Sebastiani, “Statistical phrases in automated text categorization,” *Tech. Rep. IEI-B4-07-2000, Pisa, Italy*, pp. 1–18, 2000.

[22] A. Lehmam, “Essential summarizer: innovative automatic text summarization software in twenty languages,” in *Adaptivity, Personalization and Fusion of Heterogeneous Information*, 2010, no. Lehmam, pp. 216–217.

[23] S. Shehata, F. Karray, and M. S. Kamel, “An efficient concept-based retrieval model for enhancing text retrieval quality,” *Knowl. Inf. Syst.*, vol. 35, no. 2, pp. 411–434, 2013.

[24] D. Burley, “Information Visualization As A Knowledge Integration Tool,” *J. Knowl. Manag. Pract.*, vol. 11, no. 4, pp. 1–8, 2010.

[25] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant, “Discovering interesting usage patterns in text collections: integrating text mining with visualization,” in *Main*, 2007, pp. 213–221.

[26] S. Dhawan, K. Singh, and V. Khanchi, “A Framework for Polarity Classification and Emotion Mining from Text,” *Int. J. Eng. ...*, 2014.

[27] M. N. M. Shelke, “Approaches of Emotion Detection from Text,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 2, no. 2, pp. 123–128, 2014.

[28] N. Kiyavitskaya, N. Zeni, L. Mich, J. R. Cordy, and J. Mylopoulos, “Text Mining Through Semi Automatic Semantic Annotation,” in *Practical Aspects of Knowledge Management*, Springer, 2006, pp. 143–154.

[29] H. Tan and P. Lambrix, “Selecting an ontology for biomedical text mining,” in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 2009, no. June, pp. 55–62.

[30] C. Zhai, “A cross-collection mixture model for comparative text mining,” in *In Proceedings of KDD ’04*, 2004, pp. 743–748.

[31] K. Sathiyakumari, V. Preamsudha, and G. Manimekalai, “Unsupervised Approach for Document Clustering Using Modified Fuzzy C mean Algorithm,” *Int. J. Comput. Organ. Trends*, vol. 1, no. 3, pp. 10–14, 2011.

[32] S. T. Li and F. C. Tsai, “A fuzzy conceptualization model for text mining with application in opinion polarity classification,” *Knowledge-Based Syst.*, vol. 39, pp. 23–33, 2013.

[33] G. Mamakis, a. G. Malamos, and J. a. Ware, “An alternative approach for statistical single-label document classification of newspaper articles,” *Journal of Information Science*, vol. 37, no. 3. pp. 293–303, 2011.

[34] a. Joorabchi and a. E. Mahdi, “An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata,” *Journal of Information Science*, vol. 37, no. 5. pp. 499–514, 2011.

[35] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Gonçalves, and W. Meira, “Word co-occurrence features for text classification,” *Inf. Syst.*, vol. 36, no. 5, pp. 843–858, 2011.

[36] J. R. Nolan, “Computer systems that learn: An empirical study of the effect of noise on the performance of three classification methods,” *Expert Syst. Appl.*, vol. 23, no. 1, pp. 39–47, 2002.

[37] P. Jörgensen, “Incorporating context in text analysis by interactive activation with competition artificial neural networks,” *Inf. Process. Manag.*, vol. 41, no. 5, pp. 1081–1099, 2005.

[38] C. Chen, F. S. C. Tseng, and T. Liang, “Data & Knowledge Engineering An integration of WordNet and fuzzy association rule mining for multi-label document clustering,” *Datak*, vol. 69, no. 11, pp. 1208–1226, 2010.

[39] C. Fellbaum, “WordNet,” in *Theory and Applications of Ontology: Computer Applications*, Springer Netherlands, 2010, pp. 231–243.

[40] M. U. Devi and G. M. Gandhi, “Wordnet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports Domain,” *J. Comput. Sci.*, vol. 11, no. 2, pp. 361–371, Feb. 2015.

[41] H. Costa, H. Gonçalo Oliveira, and P. Gomes, “Automatic Extraction and Validation of Lexical Ontologies from text,” no. September, p. 124, 2010.

[42] م. زاهدی, ”طراحی و پیاده سازی آنتولوژی دروس رشته ی کامپیوتر,“ ۱۳۸۸ and م. امیری.

[43] پ. دوراندیش, ”بررسی آنتولوژي در وب معنایی ,*ماهنامه وب* vol. 122, pp. 84–87, 1389.

[44] م. افتخاری, ”یک روش جدید برای انتخاب ویژگی مبتنی بر منطق فازی and ح. نصرتي ناهوك ,”*سیستم های هوشمند در مهندسی برق*, vol. 1, pp. 71–84, 1392.

[45] M. Ykhlef and S. Alqahtani, “A survey of graphical query languages for XML data,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 23, no. 2, pp. 59–70, 2011.

[46] M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi, E. Fekri, M. Monshizadeh, and S. M. Assi, “Semi Automatic Development Of FarsNet: The Persian Wordnet,” *Proc. 5th Glob. WordNet Conf.*, vol. 358, 2010.

[47] M. Montazery and H. Faili, “Automatic Persian WordNet Construction,” *Proc. 23rd Int. Conf. Comput. Linguist. (COLING 2010)*, no. August, pp. 846–850, 2010.

پیوست

واژه شناسی

واژ-هستان شناسی = wordnet

رابطهٔ تعمیم پذیری شامل دو دسته رابطهٔ هایپرنیم (Hyponymy) و هایپونیم (Hypernymy) است. هایپرنیم در واقع رابطه میان یک عبارت جزئی‌تر با یک عبارت کلی‌تر و نوعی‌تر است و با عبارت "IS_A" نیز بیان می‌شود به‌طور مثال در جمله «این یک سگ است»، می‌توان به‌جای «سگ» از عبارت کلی‌تر «حیوان» استفاده نمود. دو رابطه هایپونیم و هایپرنیم معکوس یکدیگر هستند و درنتیجه یک سگ یک هایپونیم از حیوان و حیوان یک هایپرنیم از سگ است [۱]

انواع روابط بین واژگان:

1. **hypernyms**: Y is a hypernym of X if every X is a (kind of) Y (canine is a hypernym of dog)

   اگر واژه x «نوعی از» واژه y باشد آنگاه واژه y یک **hypernym** برای واژه x است. (تعمیم پذیری)

2. **hyponyms**: *Y* is a hyponym of *X* if every *Y* is a (kind of) *X* (*dog* is a hyponym of *canine*)

   اگر واژه y «نوعی از» واژه x باشد آنگاه واژه y یک **hyponym** برای واژه x است. (شمول)

3. **Coordinate terms**: *Y* is a coordinate term of *X* if *X* and *Y* share a hypernym (*wolf* is a coordinate term of *dog*, and *dog* is a coordinate term of *wolf*)

اگر دو واژه x و y در hypernym مشترک باشند آنگاه این دو واژه **coordinate term** یکدیگر می‌باشند.

4. **holonym**: *Y* is a holonym of *X* if *X* is a part of *Y (building* is a holonym of *window*)

اگر واژه x «بخشی از» واژه y باشد آنگاه واژه y یک **holonym** برای واژه x است.

5. **meronym**: *Y* is a meronym of *X* if *Y* is a part of *X* (*window* is a meronym of *building*)

اگر واژه y «بخشی از» واژه x باشد آنگاه واژه y یک **meronym** برای واژه x است. (بخش پذیری یا جزواژگی)

تعریف sense[۳۷]:

معنی لغت در WordNet. هر sense از لغت در یک synset متفاوت است.

تعریف synset:

تعریف ۱: هر مفهوم معنادار در wordnet که احتمالاً توسط چندین لغت یا عبارت توصیه شده، sysnset نامیده می‌شود.[۳۸]

تعریف ۲: مجموعه‌ای از یک یا چند مترادف[۳۹]

تعریف ۳: یک مجموعه مترادف. مجموعه‌ای از لغات که در برخی متون بدون تغییر مفهوم جمله‌ای که در آن درج شده‌اند قابل جایگذاری هستند[۴۰].

[37] Meaning of a word in WordNet. Each sense of a word is in a different

[38] Each meaningful concept in WordNet, possibly described by multiple words or word phrase, is called a "synset."

[39] a set of one or more synonyms

[40] A synonym set; a set of words that are interchangeable in some context without changing the truth value of the preposition in which they are embedded

Abstract

By rapid growth of artificial intelligence science, new methods in this field have better performance and fewer errors. Natural language processing is a topic in artificial intelligence. So many articles are written about text mining, most of which are aimed at text classification and some clustering approaches have used fuzzy methods. One of tasks in the field of natural language processing and text mining is creation of lexical database WordNet. Lexical database WordNet is a set of words in synonym sets by which knowledge base is created. The database was created by Princeton university and also developed in other languages. FarsNet as a model of WordNet in Persian is implemented by laboratory of natural language processing in Shahid Beheshti university. FarsNet deals with different semantic categories for every word and in fact is an ontology. FarsNet is an introduction to Persian knowledge base.

In this thesis modified lexical database named FuzzNet is created by means of lexical database FarsNet. In fact Fuzznet is a combination of FarsNet and fuzzy theory. As regards relation between a word with respected meanings is crisp in FarsNet, the relation is defined fuzzy instead of crisp, That is each word is related to its meaning by fuzzy values. Membership value of a meaning for a word is higher if the meaning is closer to mind. The fuzzy membership value is helpful in translation as that meaning with higher membership value for the word is selected and translation for the meaning is done.

Keywords: fuzzy, FarsNet, WordNet, translation, database

Shahrood University of Technology

Computer engineering and information technology

Design and implementation of a lexical dataset named Fuzznet considering fuzzy theory for statistical machine translation

Mohammad Nazarian

Supervisor:

Dr. Morteza Zahedi

Date: **September 2015**