دانشکده مهندسی کامپیوتر

ابهام‌زدایی و ارزیابی اطلاعات استخراج شده از متن زبان طبیعی

سمینار کارشناسی ارشد

گروه هوش مصنوعی

محسن ایمانی

m_imany@comp.iust.ac.ir

استاد راهنما:

دکتر مرتضی آنالویی

آذر ۱۳۹۲

چکیده

سامانه‌های استخراج اطلاعات اغلب اطلاعات را به صورت یک رابطه به همراه مجموعه آرگومان‌هایش استخراج می‌نمایند. با اجرای این فرآیند در دامنه وسیعی از داده‌ها مثل وب، می‌توان صدها میلیون مجموعه به صورت مذکور بدست آورد که شامل میلیون‌ها رابطه متفاوت به همراه آرگومان‌هایش باشد[۱].

در این مرحله یک استخراج اطلاعات با کیفیت با دو چالش اصلی مواجه خواهد بود. مسئله اول این است که هر مفهوم ممکن است در قالب مجموعه‌ای از کلمات متفاوت در متن بیان شده باشد. یک استخراج اطلاعات با کیفیت باید از رابطه‌ها ابهام‌زدایی نموده و رابطه‌های یکسان را تشخیص دهد و همچنین با ابهام‌زدایی از آرگومان‌ها، هر آرگومان را به موجودیت متعلق به آن الحاق نماید. چالش دوم در استخراج با کیفیت اطلاعات ارزیابی است. سامانه باید بتواند میزان صحت و همچنین اهمیت اطلاعات استخراج شده خود را ارزیابی نماید.

در این سمینار ابتدا به شرح مفهومی با نام مطالعه ماشینی می‌پردازیم . سپس به طور خلاصه روش‌های متفاوت استخراج اطلاعات معرفی کرده و شرح می‌دهیم. سپس ابهام و ارزیابی را به عنوان چالش‌های موجود در اطلاعات استخراج شده از این روش‌ها، با نگاه ویژه به استخراج آزاد اطلاعات، معرفی می‌نماییم.

در بخش بعدی ابتدا به بررسی روش‌های ابهام‌زدایی از اطلاعات، شامل روش‌های ابهام‌زدایی از موجودیت‌ها، روش‌های ابهام‌زدایی از روابط و همچنین روش‌های ابهام‌زدایی هم‌زمان موجودیت‌ها و روابط خواهیم پرداخت. سپس روش‌های مواجهه با چالش ارزیابی اطلاعات را، با نگاه ویژه به استخراج آزاد اطلاعات، مورد بررسی قرار خواهیم داد.

در نهایت نیز کاربردهای استخراج آزاد اطلاعات، با فرض ارتقای کیفیت اطلاعات از طریق روش‌های ابهام‌زدایی معنایی و ارزیابی، در حوزه‌های مختلف مربوط به پردازش زبان طبیعی معرفی و بررسی می‌نماییم.

واژه‌های کلیدی:

استخراج اطلاعات، ابهام‌زدایی اطلاعات، ارزیابی اطلاعات، استخراج آزاد اطلاعات، پردازش زبان طبیعی

فهرست مطالب

فصل اول:

مقدمه

۱.مقدمه

مدت‌هاست که نوشتار جایگاه ثبت دانش و تجربه‌ی بشری است و متن از نگاه پژوهشگران حوزه استخراج اطلاعات، مجموعه‌ای غیر ساخت‌یافته از رابطه‌هایی است که میان مفاهیم و موجودیت‌های جهان واقعی وجود دارد. به این موجودیت‌ها و روابط میان آن‌ها اطلاعات گفته می‌شود، و در واقع هدف از استخراج اطلاعات تبدیل متن خام غیرساخت‌یافته به اطلاعات استخراج شده و قابل استفاده توسط ماشین می‌باشد.

۱,۱.مطالعه ماشینی

زمان زیادی از آغاز پژوهش‌ها در زمینه‌ی پردازش زبان طبیعی نمی‌گذرد. در این مدت بخش زیادی از توان پژوهش‌گران صرف حل مسائل خوش‌تعریف و البته محدود به دامنه کوچک شده است. برای نمونه مسئله تشخیص برچسب اجزای سخن[۱] را می‌توان یک رده‌بندی[۲] مشخص دانست. حتی مسائل پیچیده‌تری مثل تجزیه نحوی یا معنایی نیز به این شکل قابل تعریف هستند. روح کلی مفاهیم مورد بررسی در این پژوهش بیشتر با «فهم متن» سازگار است و منظور از فهم متن، چیزی فراتر از معنی با تعریف کلاسیک آن است. در واقع واژه معنا[۳] در پردازش زبان طبیعی کاملا چارچوب‌بندی شده است؛ یعنی ماشین موظف است که نقش معنایی یک واژه را از میان گزینه‌های از پیش تعریف شده انتخاب نماید.

مطالعه ماشینی واژه‌ای برای فهم مورد نظر ما از متن است. واضح است که این مسئله، یک مسئله بی‌ناظر خواهد بود. این تعریف عمومی با دنیای ماشین ناسازگار است و حتی اگر آن را فهم «بی‌ناظر» بنامیم، همچنان گنگ به نظر می‌رسد[۲].

بحثی که از فهم ماشینی در حوزه علوم شناختی[۴] مطرح می‌شود، در این حوزه مورد نظر نیست و منظور از فهم توسط ماشین در این حوزه، همان شکل ساخت‌یافته اطلاعات، یعنی روابط میان موجودیت‌هاست.

به این ترتیب چیزی که ما به عنوان فهم از ماشین توقع داریم، پایین‌تر از فهم انسانی خواهد بود. برای حل این مشکل، سعی می‌شود از نقاط قوت ماشین برای پوشاندن نقاط ضعف آن استفاده نمود. در مسئله مطالعه ماشینی، اگر چه انسان می‌تواند با دقت بسیار بالایی روابط میان موجودیت‌ها را استخراج نماید. اما از پردازش انبوه اطلاعات عاجز خواهد بود. در واقع آن‌چه که در استخراج روابط برای انسان به صورت دقت بسیار بالا و سرعت پایین وجود دارد، در ماشین می‌تواند به طور عکس آن نمود پیدا کند. ماشین توان تحلیل یکپارچه‌ی پیکره‌ی عظیمی از متون را دارد، هرچند که دقت استخراج روابط در یک متن از انسان پایین‌تر باشد. اما این

[1] Part of Speech tagging
[2] Classification
[3] Semantic
[4] Cognitive Scince

تحلیل یکپارچه‌ی پیکره عظیم متنی که در انسان در حد تحلیل متن چند صفحه کاهش می‌یابد. می‌تواند به عنوان یک نقطه قوت ماشین، نقطه ضعف آن یعنی دقت پایین‌تر را پوشش دهد.

۱.۱,۲استخراج اطلاعات

همانطور که گفتیم، استخراج اطلاعات به منظور تبدیل متن به اطلاعات قابل استفاده از منظر ماشین است. رویه‌ی مرسوم در پژوهش‌های دیرین مرتبط با این موضوع مبتنی بر اهداف از پیش تعیین شده بوده است. یعنی ماشین باید نوع خاصی از اطلاعات را که احتمالا به اشکال مشخصی نیز در متن بیان می‌شوند، استخراج می‌کرده. «استخراج هدفمند اطلاعات» عنوان مناسبی برای این نحوه نگاه به مسأله به نظر می‌رسد. مثلا استخراج زمان و مکان برگزاری کنفرانس‌ها را می‌توان یک استخراج هدفمند دانست که معمولا باید روی دامنه مشخصی از متن‌ها (مثل اعلان برگزاری کنفرانس‌ها) انجام گیرد.

۱.۱,۲,۱استخراج هدفمند اطلاعات

استخراج اطلاعات به منظور تبدیل متن خام و غیر ساخت‌یافته به اطلاعات قابل استفاده توسط ماشین مورد نظر قرار گرفته‌است. استخراج آزاد به تعبیر سنتی و مرسوم آن در دهه‌های پیشین، بدین صورت قلمداد می‌شد که ماشین باید نوع خاصی از اطلاعات و روابط را که احتمال به اشکال مشخصی در متن بیان می‌شود استخراج کند. در واقع می‌توان این منظر را به «استخراج هدفمند اطلاعات» تعبیر نمود.

در این منظر، استخراج اطلاعات به صورت یک مسئله با ناظر تعریف می‌شود. از روش‌های استخراج هدفمند اطلاعات می‌توان به روش‌های مبتنی بر قاعده که در سامانه‌هایی از قبیل YAGO [۳]، DBpedia [۴] از آن استفاده شده است. در این سامانه‌ها با استفاده از قواعد دست‌ساز، انبوه اطلاعات ساخت‌یافته موجود در ویکی‌پدیا و یا وردنت استخراج می‌شوند. روش دیگر استخراج هدفمند اطلاعات، استفاده از مدل‌های گرافی است. برای نمونه استخراج ویژگی‌های مقاله از میان سربرگ و ارجاع‌ها به شکل یک مسئله پیش‌بینی ساختار تعریف و حل شده است[۵]. روش دیگر استخراج اطلاعات استفاده از توابع کرنل است که برای این کار تعریف شده و مورد استفاده قرار گرفته‌اند. برای نمونه استفاده از تجزیه کم‌عمق جمله برای تشخیص رابطه اشخاص و نهادها و همچنین مکان سازمان‌ها بررسی شده‌است[۶].

۱.۱,۲,۲استخراج آزاد اطلاعات

استخراج هدفمند اطلاعات نیازمند صرف توان زیاد انسانی برای مشخص کردن محدوده و نوع دانش مورد تقاضا است و این یعنی رویه هدفمند توان مواجهه با حجم عظیم و متنوعی از اطلاعات را ندارد. وقتی از استخراج آزاد اطلاعات صحبت می‌کنیم، دقیقا در پی روشی برای استخراج روابط کاملا متنوع از انبوه متون درون وب هستیم. ورودی سامانه‌های استخراج آزاد اطلاعات حجم عظیمی از متن خام و خروجی آن، اطلاعات به شکل روابط میان موجودیت‌ها درون متن می‌باشد.

سامانه‌ی textrunner اولین سامانه‌ای بود که با معرفی پارادایم «استخراج آزاد اطلاعات» عرضه شد[۷]. این سامانه از سه بخش اساسی تشکیل شده است و می‌توان گفت یک چارچوب قابل توسعه برای استخراج انبوه

اطلاعات از متن خام را فراهم می‌آورد. این سامانه ابتدا با اعمال تعدادی قانون روی داده‌ها، برای خود تعدادی نمونه‌ی صحیح ایجاد کرده و بعد آن‌ها را یاد می‌گیرد. این روش را با نام روش یادگیری خود-ناظر نامگذاری کرده‌اند. سپس ابزار رده‌بند آموزش دیده با روش خود-ناظر برای استخراج روابط از داده‌ها استفاده می‌شود. در نهایت نیز با توجه به میزان تکرار روابط با موجودیت‌ها در مورد صحت نتایج قضاوت صورت می‌گیرد.

ایده‌ی موفق textrunner توسط سامانه‌های زیادی مورد استفاده قرار گرفت که از آن جمله می‌توان به WOE [۸] اشاره کرد که با استفاده از داده‌های ساخت‌یافته‌ای که در صفحات ویکی‌پدیا وجود دارد، داده‌های مورد نیاز برای آموزش را ایجاد می‌نماید. در واقع این سامانه داده‌های ساخت‌یافته‌ی هر صفحه را در متن صفحه جستجو می‌کند و به دنبال جملاتی می‌گردد که این اطلاعات در آن بیان شده باشد. جمله بازیابی شده برای آن رابطه به عنوان نمونه مثبت در نظر گرفته می‌شود.

سامانه‌ی Reverb نیز یکی دیگر از سامانه‌های استخراج آزاد اطلاعات می‌باشد که اساس آن بر چند قاعده ساده بنا شده است[۹]. فرآیند استخراج اطلاعات در این سامانه با پیدا کردن فعل‌ها در متن آغاز شده و سپس رابطه‌ی متناسب با هر فعل استخراج می‌شود.

سامانه SnowBall نیز یک سامانه نیمه‌نظارتی برای استخراج اطلاعات می‌باشد[۱۰]. در این روش سامانه تعدادی داده‌ی ورودی آغاز به کار کرده و سعی می‌کند الگوهای مربوط به وقوع‌های مختلف این نمونه‌ها را یاد بگیرد.

سامانه KowItAll بر خلاف SnowBall نیازی به داده‌ی ابتدایی برای شروع فرآیند استخراج اطلاعات ندارد[۱۱]. این سامانه برای شروع کار نیاز به تعدادی الگو و البته شرح داده‌ی مورد نظر برای استخراج دارد. این الگوها وابسته به زبان و البته مستقل از رابطه می‌باشد. این سامانه با استفاده از این الگوها و داده‌ی مورد نظر تعدادی عبارت تولید می‌کند و با استفاده از موتور جستجو، صفحات وب مرتبط با آن‌ها را بازیابی می‌نماید. صفحات بازیابی شده با استفاده از قانون موجود در الگو مورد تحلیل قرار گرفته و در نهایت اطلاعات از آن استخراج می‌شوند.

سامانه طراحی شده در [۱۲] نیز نظیر سامانه Textrunner روابط را از پیکره عظیمی از متون وب استخراج به صورت بی‌ناظر استخراج می‌نماید. این سامانه الگوهای مختص هر نوع رابطه را از طریق مجموعه‌ای از داده‌های آغازین دریافت نموده، سپس روابط کاندید را با استفاده از این الگوها کشف می‌نماید، و نهایتاً روابط با امتیاز بالاتر را به مجموعه نمونه‌های آغازین اضافه کرده و این کار را تا زمانی که به شرایط همگرایی برسد ادامه خواهد داد.

۱,۳.چالش‌های اطلاعات استخراج شده

علیرغم پیشرفت‌های عظیمی که در سال‌های اخیر در استخراج اطلاعات به وقوع پیوسته است، همچنان مسائلی در حوزه‌های کیفیت، قابلیت اعتماد و همچنین پیچیدگی روش‌های استخراج اطلاعات وجود دارند که سامانه‌ها و روش‌های استخراج اطلاعات را در عمل دچار چالش نموده است. مسئله تولید یک پایگاه دانش به طور عمومی شامل مراحل انتخاب منابع داده، استخراج موجودیت‌ها و روابط موجود در داده و در نهایت

یکپارچه‌سازی و مرتبط ساختن این اطلاعات جدید با دانش قبلی موجود در پایگاه دانش می‌شود. با وجود این که هر کدام از این مراحل توسط برخی از روش‌های ذکر شده در بخش گذشته پوشش داده شده‌اند، اما نتایج آن‌ها اغلب کیفیت لازم را برای استفاده‌ی کاربردی ندارند. بنابراین، هنوز هم بسیاری از سازمان‌ها به ناچار به تولید دستی پایگاه دانش خود روی می‌آورند.

این نیاز به کمک انسانی برای پرکردن خلا کیفی اطلاعات استخراج شده یکی از چالش‌های اساسی موجود بر سر راه روش‌های استخراج اطلاعات می‌باشد. در این بخش به بررسی چالش‌های موجود در اطلاعات استخراج شده، با نگاهی ویژه به استخراج آزاد اطلاعات خواهیم پرداخت.

۱.۱,۳,۱ابهام

هدف سامانه‌های استخراج اطلاعات از وب، نظیر آن‌هایی که در بخش گذشته معرفی شدند، استخراج روابط میان موجودیت‌ها در متن خام به شکل ساخت‌یافته می‌باشد. برای مثال جمله زیر می‌تواند نمونه‌ای از اطلاعات استخراج شده توسط این سامانه‌ها از متن باشد:

(is capital of, D.C., United States)

که نمونه بالا بیانگر وجود رابطه «is capital of» بین دو موجودیت «D.C.» و «United States» است. سامانه‌های استخراج اطلاعات از وب می‌توانند صدها میلیون از این دست اطلاعات که شامل میلیون‌ها عبارت مختلف می‌باشد از وب استخراج نمایند. یکی از مسائلی که به عنوان یک چالش حقیقی در این حوزه وجود دارد، این است که این سامانه‌ها اغلب اطلاعاتی را استخراج می‌کنند که موجودیت‌ها و یا روابط دنیای واقعی را با نام‌ها متفاوتی بیان کرده‌اند. برای مثال، همان نمونه بالا ممکن است در جایی دیگر به شکل زیر استخراج شود:

(is capital city of, Washington, U.S)

همان‌طور که مشاهده می‌نمایید این دو نمونه، دقیقا اطلاعات یکسانی را در قالب نام‌های متفاوت در بر دارند.

استفاده از کلمات هم‌معنی در متن رایج و متداول می‌باشد و متون وب نیز از این قاعده مستثنا نیستند. به عنوان نمونه مجموعه داده استخراج شده توسط TextRunner که حاوی دو میلیون نمونه اطلاعات استخراج شده از وب است، شامل شش نام متفاوت برای هر کدام از آرگومان‌های نمونه بالا، یعنی «United States» و «Washington D.C.» و سه نام متفاوت برای رابطه «is capital of» می‌باشد[۱]. همچنین در ۸۰ موجودیت پرکاربرد در این مجموعه داده به طور متوسط با ۲,۹ نام برای هر موجودیت استفاده شده‌اند که برخی از موجودیت‌ها با حتی تا ۱۰ نام متفاوت نیز استفاده شده‌اند. و نیز ۱۰۰ رابطه پرکاربرد در این مجموعه داده به طور متوسط ۴,۹ نام برای هر رابطه دارا هستند.

ما چالش وجود نام‌های هم‌معنی در اطلاعات استخراج شده از وب را «ابهام» در اطلاعات استخراج شده می‌نامیم و مسئله تشخیص نام‌های هم‌معنی در موجودیت‌ها و رابطه‌های این اطلاعات را به عنوان «ابهام‌زدایی از اطلاعات استخراج شده» مطرح می‌نماییم.

به طور خلاصه می‌توان گفت یکی از مسائلی که به عنوان یک چالش بزرگ در استخراج با کیفیت اطلاعات وجود دارد، این است که به طور معمول در متن زبان طبیعی، موجودیت‌ها و روابط دنیای واقعی با نام‌های متفاوتی استفاده می‌شوند، و سامانه‌های استخراج اطلاعات باید بتوانند این نام‌های متفاوت و متعدد را به یک موجودیت و رابطه دنیای واقعی ملحق کنند.

همچنین چالش دیگر در این زمینه این است که این نوع ابهام، در هر دو صورت روابط و موجودیت‌های اطلاعات استخراج شده وجود دارد. برخی روش‌ها در این حوزه تنها به جنبه‌ای از این چالش، یعنی ابهام‌زدایی در موجودیت‌ها به تنهایی و یا ابهام‌زدایی در روابط، به تنهایی، اشاره نموده‌اند. اما برخی روش‌ها به بررسی راه‌حل‌هایی جامع که بتواند هر دو جزء موجود در اطلاعات را ابهام‌زدایی کنند پرداخته‌اند.

علاوه بر این، برخی از روش‌ها برای ابهام‌زدایی در این حوزه وجود دارند که نیازمند مجموعه بزرگی از داده‌های آموزشی برای ابهام‌زدایی هستند و یا این که وابسته به دامنه‌ی خاصی از دانش بوده و تنها در آن دامنه خاص عمل می‌کنند. با توجه به تعداد بالا و همچنین حوزه وسیع اطلاعات استخراج شده توسط سامانه‌های استخراج آزاد اطلاعات، اساساً استفاده چنین روش‌هایی در حوزه استخراج آزاد اطلاعات غیرممکن خواهد بود. به دلیل همان مسئله تعداد بالا و حوزه وسیع، در استخراج اطلاعات از وب نمی‌توان از دانش خارجی و یا نمونه‌های آموزشی دست‌ساز برای هر رابطه استفاده نمود، و یا حداقل این که تولید دستی چنین مجموعه‌هایی نیازمند نیروی انسانی و هزینه بسیار زیادی می‌باشد.

۱.۱,۳,۲ارزیابی

همان‌طور که ذکر شد، استخراج آزاد اطلاعات، مسئله استخراج خودکار دانش از متن زبان طبیعی است. این مسئله به طور ذاتی یک مسئله بی‌ناظر و در غیاب مجموعه داده آموزشی دست‌ساز. این ذات عدم دخالت انسان و خودکار بودن استخراج دانش، مزیت بزرگی را به همراه خواهد داشت و آن قابلیت کشف سریع روابط، موجودیت‌ها و نمونه‌های جدید و قابلیت توسعه در فضاهای بسیار عظیم و با ابعاد زیاد مانند وب است.

اما این ویژگی یک چالش اساسی را نیز به همراه خواهد داشت. و این چالش چگونگی ارزیابی احتمال صحت و میزان ارزشمند بودن اطلاعات استخراج شده می‌باشد. در واقع در استخراج آزاد اطلاعات، ما با حجم عظیمی از اطلاعات استخراج شده از متن مواجه هستیم، که البته بخشی از دقت استخراج را هم به خاطر لزوم سریع بودن عملیات استخراج، به صورت عمدی از دست داده‌ایم. بنابراین با حجم زیادی از اطلاعات نامطمئن مواجه هستیم و نیاز به روش‌هایی برای ارزیابی میزان صحت این اطلاعات خواهیم داشت.

به عبارت دیگر بحث ارزیابی اطلاعات استخراج شده را می‌توان از دو جنبه بررسی نمود. جنبه اول میزان درستی استخراج اطلاعات توسط سامانه استخراج کننده اطلاعات. در واقع با توجه به این در استخراج آزاد اطلاعات، روش‌ها ذاتا به صورت بی‌ناظر و بدون استفاده از مجموعه داده آموزشی دست‌ساز عمل می‌کنند، چگونگی ارزیابی صحت و درستی اطلاعات استخراج شده توسط سامانه استخراج اطلاعات یکی از چالش‌های این حوزه خواهد بود.

ارزیابی اطلاعات از جنبه دوم، بررسی میزان صحت و ارزشمندی اطلاعات است. در واقع با توجه به این که استخراج اطلاعات در پیکره‌ی عظیمی از متن‌ها مانند وب، اطلاعات را استخراج می‌نمایند، این که اطلاعات موجود در متن سند به چه میزان ارزشمند و همچنین صحیح خواهد بود، و چگونگی تعریف معیاری برای میزان این صحت اطلاعات، چالش دیگری در حوزه ارزیابی اطلاعات خواهد بود.

در واقع سامانه استخراج کننده اطلاعات، با توجه به این که حجم بسیار عظیمی از اطلاعات را از سندهای متعدد و ناهمگون استخراج خواهد کرد، باید بتواند اطلاعات ناصحیح و همچنین اطلاعات غیرمفید را از این مجموعه عظیم از اطلاعات تشخیص داده و این مجموعه را تصفیه نماید.

همین حجم عظیم اطلاعات استخراج شده و تعدد بالای سندها و ناهمگونی آن‌ها در پیکره‌های عظیمی مانند وب، خود می‌تواند منشاء طراحی روش‌هایی برای ارزیابی بدون ناظر اطلاعات باشد. در واقع ایده اصلی این روش‌ها این است که اطلاعاتی که در این مجموعه عظیم و ناهمگون میزان تکرار بالاتری داشته باشند، می‌توان نتیجه گرفت که میزان احتمال صحت این اطلاعات نیز بالاتر خواهد بود.

فصل دوم:

روش‌های ابهام‌زدایی و ارزیابی اطلاعات

۲.روش‌های ابهام‌زدایی و ارزیابی اطلاعات

تا به این‌جا به بررسی اجمالی روش‌های استخراج اطلاعات از متن پرداختیم و سپس به طرح مسئله پیرامون دو چالش اساسی در این حوزه، یعنی ابهام در اطلاعات و همچنین ارزیابی اطلاعات، پرداختیم.

در این فصل به بررسی روش‌ها و راهکارها برای هر یک از این چالش‌ها خواهیم پرداخت.

۲,۱. ابهام‌زدایی اطلاعات استخراج شده

تا به این‌جا به بررسی اجمالی روش‌های استخراج اطلاعات از متن پرداختیم و سپس به طرح مسئله پیرامون دو چالش اساسی در این حوزه، یعنی ابهام در اطلاعات و همچنین ارزیابی اطلاعات، پرداختیم.

در این بخش به بررسی روش‌های مواجهه و حل چالش اول، یعنی روش‌های ابهام‌زدایی از اطلاعات استخراج شده، با نگاه ویژه به استخراج آزاد اطلاعات، خواهیم پرداخت.

۲,۱,۱.چالش‌ها

به طور کلی می‌توان گفت در راه حل‌های ارائه شده برای این مسئله، برخی روش‌ها تنها به جنبه‌ای از مسئله، یعنی ابهام‌زدایی از موجودیت‌ها و یا روابط، به تنهایی، پرداخته‌اند. اما برخی روش‌ها نیز راه‌حل‌های جامعی را برای ابهام‌زدایی هر دو جزء در اطلاعات ارائه داده‌اند.

همچنین برخی روش‌های با ناظر در این حوزه نیازمند نمونه‌های آموزشی اولیه هستند و یا از دانش حوزه موضوعی خاصی برای ابهام‌زدایی در آن حوزه معنایی استفاده می‌کنند. با توجه به این که رویکرد ما در این مسئله اطلاعات استخراج شده از وب می‌باشد. در این مسئله با چالش‌هایی مواجه خواهیم بود.

اولین چالشی که در مسئله ابهام‌زدایی از اطلاعات استخراج شده از وب می‌توان به آن اشاره نمود. عدم کارایی روش‌های با ناظر است. زیرا تهیه نمونه‌های آموزشی که شامل همه الگوهای موجودیت‌ها و روابط میان آن‌ها در سطح وب باشد، عملاً ممکن نخواهد بود. در واقع مجموعه‌ی روابط میان موجودیت‌ها در دنیای حقیقی نیز یک مجموعه بی‌انتها می‌باشد. در استخراج آزاد اطلاعات نیز، ادعا بر این است که مجموعه روابط کشف شده غیر قابل شمارش بوده و به مرور زمان ممکن است به آن اضافه شود.و متن خام بهره می‌برند.

برخی از روش‌های ابهام‌زدایی، برای حل این مسئله از ویژگی‌های مضمون و متن استفاده نموده‌اند. با توجه به این که هدف از استخراج اطلاعات از وب، تحلیل یکپارچه حجم عظیمی از داده‌های وب می‌باشد، بنابراین ترجیح ما بر این است که تنها با استفاده از خود اطلاعات بتوانیم ابهام‌زدایی را انجام داده و مجبور به ذخیره داده‌ی اضافی همراه با اطلاعات نباشیم.

علاوه بر این با توجه به این که اصولاً استخراج آزاد اطلاعات، مستقل از دامنه موضوعی متن می‌باشد، بنابراین استفاده از منابع دانش خارجی که مبتنی بر دامنه‌های موضوعی خاصی هستند برای ابهام‌زدایی از این اطلاعات راه‌گشا نخواهد بود.

یک سامانه ابهام‌زدایی از اطلاعات باید به سوالات اساسی زیر پاسخ دهد:

- آیا این امکان وجود دارد که عبارات موجود در مجموعه‌ی بزرگی از اطلاعات استخراج شده را بدون استفاده از دانش خارجی مربوط به زمینه موضوعی، داده آموزشی دست‌ساز و یا سایر منابع خارجی که در استخراج اطلاعات از قابل استفاده نیست، به صورت موثر در مجموعه‌های هم‌معنی خوشه‌بندی کرد؟
- چگونه می‌توان ابهام‌زدایی معنایی را به مجموعه داده‌های عظیم و با ابعاد زیاد نظیر وب تعمیم داد؟
- چگونه می‌توانیم ابهام‌زدایی معنایی را بدون نظارت و استفاده از دانش انسانی انجام دهیم، و آیا این کار سودی برای ما دارد؟
- آیا سامانه‌های ابهام‌زدایی از اطلاعات، می‌توانند کلمات دارای تعدد معانی را که در متون مختلف با معانی متفاوت می‌آیند به درستی مدیریت کنند؟

۲,۱,۲.روش‌های ابهام‌زدایی موجودیت‌ها

ابهام‌زدایی موجودیت‌های اطلاعات، مسئله‌ای بسیار شبیه به مسئله معروف ابهام‌زدایی موجودیت‌های بین سندی[5] است؛ در واقع هدف از هر دو مسئله خوشه‌بندی موجودیت‌های اسمی ظاهر شده در مجموعه‌ای از اسناد به خوشه‌های هم‌مرجع است

آلیس[۱۳]، یک عامل یادگیرنده مادام‌العمر است که هدف آن تولید یک نظریه – مجموعه‌ای از مفاهیم، رویدادها و تعمیم آن‌ها به نحوی که دامنه‌ی خاصی را توصیف کند- به صورت مستقیم از متن خام، به صورت تکراری و سلسله مراتبی است. آلیس کار خود را با یک پیکره مربوط به دامنه‌ی خاص موضوعی، منبعی از دانش زمینه از همان دامنه، و یک استراتژی کنترل آغاز کرده و شروع به جستجو، به روز رسانی و تصفیه یک نظریه در این زمینه می‌نماید. این سامانه، حقایق مربوط به موجودیت‌ها در دامنه موضوعی را با استفاده از سامانه TextRunner از متن خام مرتبط با دامنه‌ی موضوع استخراج می‌نماید. سپس سامانه هر یک از موجودیت‌ها را در پایگاه دانش خود جستجو می‌نماید در برخی از موارد، آلیس ممکن است این اطلاعات را در پایگاه دانش خود قبلا داشته باشد، در غیر این صورت یادگیرنده‌های آلیس، کلاس‌ها و زیرکلاس‌های عضویت را برای موجودیت جدید تشخیص می‌دهند. سپس سیستم به طور خودکار اطلاعاتی که توسط TextRunner استخراج شده است را با استفاده از خصیصه‌های کلاس عضویت در پایگاه دانش، و همچنین تعداد دفعات مشاهده این نام در واژه‌های موجودیت‌ها خوشه‌بندی می‌نماید. مشکل اساسی روشی که آلیس برای ابهام‌زدایی اطلاعات استخراج شده توسط TextRunner از آن بهره می‌برد، استفاده سامانه از یک منبع دانش مرتبط با

[5] Cross-document Entity Resolution

دامنه موضوعی می‌باشد. این استفاده از منبع دانش خارجی باعث می‌شود روش مورد استفاده از حالت مستقل از دامنه خارج شود.

[۱۴] روشی را ارائه داده است که در آن نام افراد از سندها و ایمیل‌ها با استفاده از یک خوشه‌بندی تجمعی و استفاده از تابع تشابه ابتکاری ابهام‌زدایی شوند. این موجودیت‌ها را با استفاده از معیار تشابهی بر مبنای متنی که در آن ظاهر شده‌اند در گروه‌های هم‌سان خوشه‌بندی می‌کند. این روش سه مولفه‌ی اساسی دارد:

۱) شناسایی خصیصه: خصیصه‌هایی که در این روش از آن‌ها استفاده شدند شامل تک‌واژه‌ها[۶]، دو واژه‌ها[۷] و هم‌آیی نام‌هاست.

۲) نگهداری متن زمینه برای هر موجودیت: در این مولفه متنی که نام‌ها در آن ظاهر شده‌اند به شکل برداری از خصیصه‌های ذکر شده تبدیل و نگهداری می‌شوند. همچنین برای کاهش ابعاد از روش تجزیه به مقادیر منحصر به فرد[۸] استفاده شده است.

۳) خوشه‌بندی: در این بخش با استفاده از مقایسه مشابهت بین بردارهای ساخته شده از متن زمینه، خوشه‌بندی نام‌ها با استفاده از الگوریتم خوشه‌بندی سلسله‌مراتبی تجمعی صورت می‌پذیرد.

[۱۵] با استفاده از روش برآیند و بیشینه‌سازی[۹] توسط یک مدل گرافی و پایگاه‌های داده مربوط به اسامی مستعار، القاب و عناوین متداول، به سامانه‌ای با دقت بالا برای ابهام‌زدایی بین سندی موجودیت‌ها دست‌یافته است. مدلی که برای این تابع تشابه ابتکاری استفاده می‌شود دارای سه مولفه است:

۱) مولفه احتمال توزیع‌شده‌ای که نشان می‌دهد موجودیت‌ها چگونه در سندها توزیع شده‌اند و همچنین هم‌آیی موجودیت‌ها در سندها را نیز منعکس می‌کند.

۲) مولفه‌ای که شامل تعداد موجودیت‌های ظاهر شده در یک سند و تعداد دفعاتی است که هر موجودیت در سند ذکر شده است.

۳) مولفه‌ای که نشان‌دهنده‌ی میزان تشابه ظاهری بین دو نام می‌باشد.

یکی از مشکلات این روش استفاده از پایگاه داده‌ای از نام‌ها می‌باشد که باعث می‌شود سامانه وابسته به دامنه موضوعی شود.

[۱۶] نیز روشی بی‌ناظر حاصل از ادغام خصیصه‌های استخراج شده به صورت خودکار و بردار ترم‌ها که شامل اسامی خاص در متن می‌شود، برای خوشه‌بندی نام‌های مبهم در وب ارائه داده است. در این روش از معیار تشابه کوسینوسی[۱۷] و همچنین الگوریتم خوشه‌بندی سلسله‌مراتبی تجمعی بهره گرفته شده است. در واقع

---

[6] Unigrams
[7] Bigrams
[8] Singular Value Decomposition(SVD)
[9] Expectation-Maximization

در این روش نیز سندهایی که نام‌های موجودیت در آن ظاهر شده‌اند به صورت برداری از خصیصه‌های استخراج شده به صورت خودکار در نظر گرفته شده و سپس با استفاده از الگوریتم خوشه‌بندی سلسله‌مراتبی تجمعی به صورت از پایین به بالا خوشه‌بندی می‌شوند. برای استخراج خصیصه‌ها و تولید بردارهای خصیصه برای هر سند، از معیارهای زیر می‌توان استفاده نمود:

۱) صورت ظاهری کلمات و یا اسامی خاص موجود در متن

۲) کلماتی که ارتباط بیشتری با نام‌ها دارند. (با استفاده از معیارهایی نظیر tf-idf و mi)

۳) استفاده از خصیصه‌های پایه زندگینامه‌ای: برای ابهام‌زدایی از موجودیت‌ها، در صورت وجود در متن می‌توان از اطلاعاتی نظیر تاریخ تولد، شغل و ... استفاده نمود.

۴) استفاده از خصیصه‌های زندگینامه‌ای تعمیم‌یافته: این خصیصه‌ها دقیقاً مانند خصیصه‌های دسته قبل می‌باشند با این تفاوت که در این روش، به این نوع خصیصه‌ها با توجه به اهمیت اطلاعاتی بیشتری که دارند، وزن بیشتری اختصاص یافته است. مثلا اگر در متن ۱۹۸۹ به عنوان سال تولد تشخیص داده شود، وزن بیشتری از کلمه ۱۹۸۹ ای دارد که در متن به عنوان واژه ساده تشخیص داده می‌شود.

در [۱۸] نیز روشی برای ابهام‌زدایی معنایی و تولید شبکه واژگانی از اطلاعات استخراج شده به صورت بی‌ناظر ارائه شده است. البته در این روش ابهام‌زدایی معنایی تنها برای موجودیت‌ها صورت می‌گیرد. این روش ابتدا موجودیت‌های استخراج شده را با استفاده الگوریتم لسک[10][۱۹] ابهام‌زدایی معنایی می‌نماید. لسک یک الگوریتم بی‌ناظر در ابهام‌زدایی معنایی کلمات است که بر مبنای کلمات مشترک بین متنی که کلمه در آن آمده و توضیحات موجود در دسته‌های معنایی کلمه در واژه‌نامه کار می‌کند. سپس با استفاده از شبکه عصبی فازی آرت[11][۲۰] خوشه‌های موجودیت‌ها را که با روابط استخراج شده با هم مرتبط شده‌اند، به شکل یک شبکه معنایی تولید می‌کند. این روش دو نقطه ضعف عمده دارد، اول این که ابهام‌زدایی معنایی تنها در مورد موجودیت‌ها صورت می‌گیرد و راهکاری برای روابط ارائه نمی‌دهد. و دوم این که الگوریتم لسک برای ابهام‌زدایی معنایی نیاز به محتوای متنی سندها دارد و با در اختیار داشتن اطلاعات به تنهایی قادر به ابهام‌زدایی نخواهد بود.

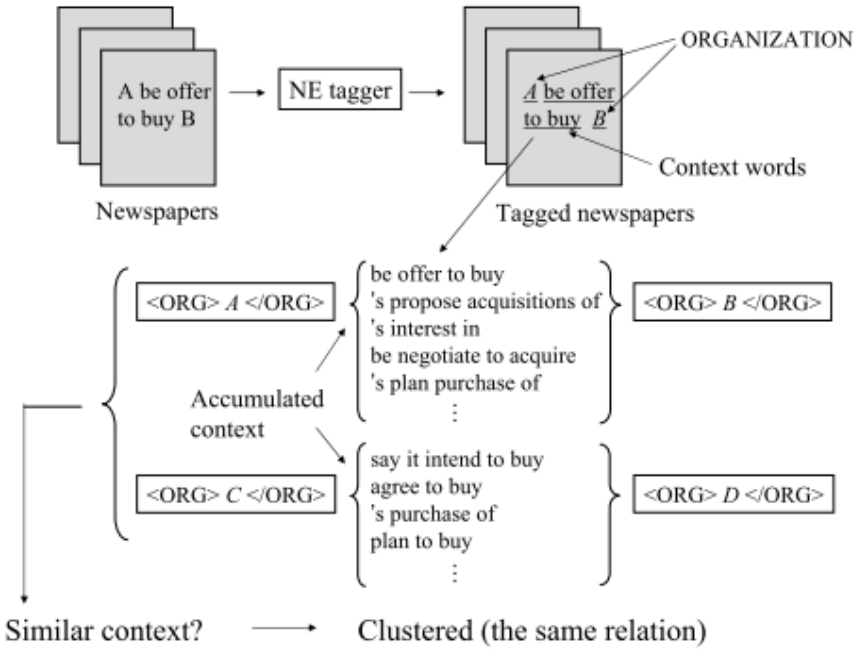
[۲۱] یک راه حل سه مرحله‌ای برای مسئله ارائه نموده است. در مرحله اول ماژول تجزیه نحوی، ساختار نحوی جمله‌ها را استخراج نموده و مسیرهای موجود از نام‌های متن به مراجع احتمالی آن‌ها را مشخص می‌سازد. چالش اساسی در این مرحله، عدم اطمینان نسبت به صحت عمل تجزیه‌گر نحوی می‌باشد و همچنین این که تجزیه نحوی نیاز به داده‌ی آموزشی ساختار درختی جمله خواهد داشت. در مرحله دوم یک ماژول معنایی سازگاری بین عناوین متن و نام‌های داخل آن را بررسی می‌نماید. نهایتاً، بعد از عملیات جداسازی نحوی و

[10] Lesk
[11] Fuzzy ART

معنایی، از بین مراجع احتمالی باقی‌مانده در متن، موجودیتی برای الحاق انتخاب می‌شود که کمترین فاصله درختی را با نام مورد نظر داشته باشد.

[۲۲] نیز روشی بی‌ناظر برای ابهام‌زدایی از موجودیت‌ها با استفاده از یک مدل گرافی ادغام شده با خصیصه‌های برجسته محلی و خصیصه‌های موجودیت کلّی ارائه نموده است. این روش نیز خصیصه‌های مورد استفاده برای خوشه‌بندی و مقایسه بین موجودیت‌ها را به صورت خودکار استخراج می‌نماید.

در [۲۳] روشی بی‌ناظر برای ابهام‌زدایی معنایی از جفت موجودیت‌ها بر مبنای روابطی که این جفت موجودیت در اطلاعات با آن ظاهر شده‌اند ارائه شده است. در روش مذکور این‌گونه فرض شده است که جفت موجودیت‌هایی که با روابط مشابهی بیان شده‌اند می‌توانند در یک خوشه معنایی قرار گیرند و این خوشه معنایی می‌تواند نمونه‌هایی از همان روابط باشد. در این روش هر رابطه بیان‌گر یک خوشه از جفت موجودیت‌ها می‌باشد. در این روش روابط با تعداد تکرار بسیار کم در یک پیکره عظیم متنی، مثلاً یک یا دو بار تکرار، نادیده گرفته می‌شوند. روال کار در این روش به این صورت است که ابتدا موجودیت‌ها برچسب خورده می‌شوند و سپس موجودیت‌هایی که در اطلاعات با هم آمده‌اند در قالب جفت موجودیت‌ها جمع‌آوری می‌شوند. سپس تشابه میان روابطی که این جفت موجودیت‌ها با هم دارند محاسبه می‌گردد و بر مبنای این تشابه، جفت موجودیت‌ها خوشه‌بندی می‌شوند. سپس هر خوشه با نام رابطه‌ای که بیشترین تکرار را برای جفت موجودیت‌ها داشته است برچسب‌گذاری می‌گردد. شکل ۱ مثالی را از این روش نشان می‌دهد.



شکل ۱: مثالی از روش ابهام‌زدایی از جفت موجودیت‌ها [۲۳]

در این مثال جفت موجودیت (Organization A, Organization B) و (Organization C, Organization D) در مجموعه اطلاعات با روابطی که با آن ظاهر شده‌اند استخراج می‌شوند. سپس تشابه بین این روابط محاسبه شده و این میزان تشابه اگر محدودیت‌ها را برآورده کند، دو جفت موجودیت فوق در یک خوشه قرار می‌گیرند. مشکل عمده‌ای که این روش دارد، این است که هر رابطه بیان‌گر یک خوشه می‌باشد و هر جفت از موجودیت‌ها تنها می‌توانند در یک خوشه حضور داشته باشند، یعنی هر جفت موجودیت تنها یک نوع رابطه می‌تواند داشته باشد که بعضاً در دنیای واقعی به این صورت نیست.

[۲۴] روشی برای خوشه‌بندی دوتایی‌های موجودیت‌ها با قابلیت هم‌پوشانی در خوشه‌ها ارائه داده است. در این روش از کلماتی که همراه با موجودیت‌ها در جمله آمده‌اند برای ساخت یک الگو استفاده می‌شود. سپس موجودیت‌هایی که در جملاتی با الگوی یکسان آمده باشند به عنوان موجودیت‌های همسان خوشه‌بندی می‌شوند. مشکل اصلی این سامانه استفاده از متن خام جملات است برای تشابه‌یابی می‌باشد.

[۲۵] روشی را برای اتصال موجودیت‌های استخراج شده توسط روش‌های استخراج اطلاعات به موجودیت مربوط به آن در ویکی‌پدیا ارائه داده است. برای مثال آرگومان Ney York در نمونه اطلاعات (New York, Acquired, Pined) باید به مقاله مربوط به تیم بیس‌بال New York Yankees متصل شود. این روش برای هر آرگومان در اطلاعات استخراج شده، مهم‌ترین موجودیت ویکی‌پدیا را که شرایط تطابق رشته‌ای را داشته باشد، پیدا می‌کند. این روش معیار مهم بودن موجودیت‌های ویکی‌پدیا را از روی لینک‌های ورودی آن، یعنی تعداد صفحات ویکی‌پدیا که به این صفحه این موجودیت متصل شده‌اند، محاسبه می‌نماید. برای مثال، در مورد نمونه قبلی موجودیت‌های کاندید برای آرگومان "New York" موجودیت‌هایی نظیر موارد موجود در جدول ۱ خواهند بود.

جدول ۱: تعداد لینک‌های ورودی به صفحات مربوط به موجودیت‌های کاندید ویکی‌پدیا برای آرگومان "New York"

موجودیت	تعداد لینک‌های ورودی
New York (State)	۹۲۲۵۳
New York (City)	۸۷۹۷۴
New York Yankees	۸۶۴۷
New York University	۷۹۸۳

بعد از به دست آوردن لیست موجودیت‌های کاندید، در مرحله تطبیق محتوا با استفاده از معیار شباهت کوسینوسی[۱۷] فاصله معنایی بین نام مورد نظر و هر کدام از مقالات کاندید ویکی‌پدیا محاسبه می‌شود. سپس این روش با استفاده از دو معیار امتیاز لینک که حاصل از ضرب معیارهای تطبیق رشته، اهمیت صفحه و همچنین تطبیق محتواست و همچنین معیار ابهام لینک که حاصل تقسیم امتیاز دومین پرامتیازترین لینک بر دومین کم‌امتیازترین لینک، موجودیتی که بیشترین امتیاز لینک و کم‌ترین ابهام لینک را دارد به عنوان موجودیت مربوط به نام مورد نظر انتخاب می‌شود.

۲,۱,۳. روش‌های ابهام‌زدایی روابط

ابهام‌زدایی معنایی از روابط اغلب مشابه مسئله «کشف نقل مضمون[۱۲]» می‌باشد. سامانه‌های توسعه داده شده در این زمینه از قبیل [۲۶]، [۲۷] و [۲۸] برای حل این مسئله از پیکره‌های موازی و هم‌تراز شده استفاده، بدین صورت که نقل‌های متفاوت از یک متن یا گزارش‌های خبری متفاوت از یک رویداد به صورت موازی در یک پیکره هم‌تراز شده باشند، برای تشخیص نقل مضمون‌ها استفاده نموده‌اند.

در [۲۹] از مجموعه داده‌ی برچسب خورده به صورت دستی برای آموزش با ناظر مدل نقل مضمون‌ها استفاده نموده است.

پاسکال[۳۰] مسئله تشخیص استلزام متنی[۱۳] را برای تشخیص این که چه زمانی یک جمله، جمله‌ی دیگر را نتیجه می‌دهد، پیشنهاد نموده است. این مسئله به صورت تشخیص استلزام را با استفاده از مجموعه داده آموزشی برچسب‌خورده به صورت دستی تعریف شده است و بسیاری از پژوهش‌گران راه حل‌هایی را برای این چالش ارائه داده‌اند. مشکل این راه حل‌ها ذات با ناظر بودن آن‌ها و در نتیجه استفاده از منابع دست‌ساز می‌باشد که قابلیت تعمیم آن را در دامنه‌های وسیع موضوعی و متنی و پیکره‌های عظیمی از جمله وب را ندارند.

برای مسئله کشف نقل مضمون، سامانه‌های متعددی نیز به صورت بی‌ناظر پیشنهاد شده‌اند که تمرکز آن‌ها روی خوشه‌بندی معنایی روابط بر مبنای پیکره‌های متنی می‌باشد.

[۳۱] از یک روش ابتکاری محاسبه معیار شباهت برای خوشه‌بندی رابطه‌ها استفاده نموده است. این روش در قالب یک پروسه چهار مرحله‌ای راه حلی برا مسئله ابهام‌زدایی از روابط ارائه نموده است. در مرحله ابتدایی جفت موجودیت‌های اسمی به همراه زمینه متنی آن‌ها استخراج می‌شوند. در مرحله دوم برای هر جفت موجودیت اسمی، کلمات کلیدی از متن زمینه آن‌ها استخراج می‌شود. این کلمات کلیدی در واقع نشان‌دهنده‌ی عبارتی خواهند بود که از آن استخراج شده‌اند. در مرحله سوم، عبارت‌هایی که دارای کلمه کلیدی یکسان هستند در یک گروه قرار می‌گیرند. و در نهایت، در آخرین مرحله سعی می‌شود تا عباراتی که یک رابطه را با کلمات کلیدی متفاوت نشان داده می‌شوند نیز در یک گروه، خوشه‌بندی شوند. برای این کار از جفت موجودیت‌های اسمی مربوط به هر عبارت کمک گرفته شده است. به این ترتیب که عبارت‌های با کلمات کلیدی متفاوت، که جفت موجودیت اسمی آن‌ها مشابه هستند، احتمالاً یک رابطه را بیان می‌نمایند.

[۳۲] یک خوشه‌بندی با استفاده از روش ابتکاری دسته‌بندی الگوهایی از رابطه‌ها که برای استخراج نمونه‌های رابطه استفاده می‌شوند، ارائه داده است. این الگوریتم ابتدا الگوهایی که با هم در یک کلمه قلّاب[۱۴] اشتراک

[12] Paraphrase discovery
[13] Textual Entailment
[14] Hook word

داشته باشند در یک خوشه قرار می‌دهد. سپس خوشه‌های حاصل را به صورت دو به دو تا رسیدن به ساختار نهایی به همین ترتیب ادغام می‌نماید. نهایتاً، الگوریتم خوشه‌هایی از الگوهای روابط هم‌معنا به عنوان حاصل ایجاد خواهد کرد که نمونه روابط ایجاد شده با این الگوها نیز هم‌معنا خواهند بود.

۲,۱,۴. روش‌های ابهام‌زدایی هم‌زمان موجودیت‌ها و روابط

۲,۱,۴,۱. سامانه RESOLVER

Resolver یک سامانه بی‌ناظر و مستقل از دامنه برای ابهام‌زدایی از هر دو جزء اطلاعات، یعنی هم موجودیت‌ها و هم روابط، می‌باشد[۳۳]. این سامانه نام‌های هم‌معنی را با استفاده از یک مدل احتمالاتی که با تشابه رشته‌ای[۱۵] و تشابه میان اطلاعاتی که این نام‌ها را در بر دارند، ساخته شده است، خوشه‌بندی می‌نماید.

۲,۱,۴,۱,۱.مدل‌های احتمالاتی Resolver

این سامانه از مدلی برای تشخیص هم‌معنی‌ها به صورت بی‌ناظر و در غیاب مجموعه داده آموزشی استفاده می‌کند. برای این کار، Resolver از دو منبع بهره می‌گیرد: تشابه بین نام‌ها[۱۶] و همچنین تشابه بین اطلاعاتی که این نام‌ها در آن‌ها ظاهر شده‌اند، که منبع دوم بعضاً با نام تشابه توزیع شده نیز شناخته می‌شود.[۳۴]

بسیاری از موجودیت‌ها با نام‌ها متفاوتی ظاهر می‌شوند که این نام‌ها اغلب یکی از آن‌ها به صورت هم‌خانواده، زیررشته و یا مخفف دیگری، و یا شکل متفاوتی از نام دیگر می‌باشد. بنابراین تشابه بین رشته‌ای می‌تواند منبع مهمی برای تشخیص دو واژه هم‌معنی باشد. Resolver نیز از یک مدل احتمالاتی تشابه بین رشته‌ای[۱۷] استفاده می‌کند. این مدل یک تابع تشابه به صورت تابع $sim(s1,s2): String \times String \rightarrow [0,1]$ فرض می‌کند که با استفاده از این تابع، مدل احتمال این که دو رشته s1 و s2 هم‌معنی باشند به صورت نشان داده شده در رابطه ۱ تعریف می‌شود.

$$P\left(R_{i,j}^{t} \mid sim(s_1, s_2)\right) = \frac{\alpha * sim(s_1, s_2) + 1}{\alpha + \beta}$$

رابطه1

α و β در این مدل پارامترهای مسئله هستند و طوری انتخاب می‌شوند که احتمال نهایی هیچ‌گاه برابر صفر یا یک نباشد. در این سامانه برای موجودیت‌ها از معیار تشابه میان رشته‌ای مونگ-الکان[۱۸][۳۵] و برای روابط از معیار فاصله میان رشته‌ای لونشتین[۱۹] استفاده شده است[۳۶].

[15] String similarity
[16] Edit distance
[17] Probabilistic String Similarity Model(SSM)
[18] Monge-Elkan
[19] Leveneshtein

مدل دیگری که این سامانه از آن بهره برده است، مدل احتمالاتی ویژگی مشترک استخراج شده[20] می‌باشد. این مدل احتمال این که دو رشته هم‌مرجع باشند را بر مبنای تشابه اطلاعات استخراج شده‌ای که در آن ظاهر شده‌اند محاسبه می‌نماید. برای مثال اگر دو نمونه از اطلاعات را به شکل (invented, Newton, calculus) و (invented, Leibniz, calculus) در نظر بگیریم، دو واژه Newton و Leibniz در این نمونه‌ها دارای مضمون مشترک در نظر گرفته می‌شوند.

به بیانی دیگر، فرض کنید که دوتایی رشته‌های (r, s) به عنوان ویژگی‌های موجودیت o قلمداد می‌شوند اگر اطلاعاتی به شکل $(r, o, s) \in D$ و یا $(r, s, o) \in D$ وجود داشته باشد. همچنین دوتایی (s1, s2) نمونه‌ای از رابطه‌ی r هستند اگر اطلاعاتی به صورت $(r, s1, s2) \in D$ وجود داشته باشد. به همین صورت ویژگی p = (r, s) متعلق به موجودیت o بوده و همچنین نمونه i=(s1, s2) نیز متعلق به رابطه r می‌باشد. مدل ویژگی مشترک استخراج شده احتمال این که دو رشته هم‌مرجع باشند را بر مبنای این که چند ویژگی (یا نمونه) مشترک با هم دارند حساب می‌کند.

برای مثال، دو رشته Mars و Red planet را در نظر بگیرید که در مجموعه اطلاعات استخراج شده در [۱] به ترتیب ۶۵۹ و ۲۶ بار ظاهر شده‌اند. در این مجموعه از اطلاعات استخراج شده، این دو رشته ۴ ویژگی مشترک دارند. برای مثال دو نمونه (lacks, Mars, ozone layer) و (lacks, Red planet, ozone layer) هر دو در مجموعه داده وجود دارند. این مدل احتمال این که دو رشته Mars و Red planet به موجودیت یکسانی مربوط باشند را، بعد از مشاهده k ویژگی مشترک از n1 ویژگی مشاهده شده در Marsو n2 ویژگی مشاهد شده در Red planet مشخص می‌نماید.

مدل ویژگی مشترک استخراج شده را می‌توان شبیه به مدل توپ و ظرف در نظر گرفت[۳۷]. برای هر رشته si، تعداد مشخص Pi ویژگی از این رشته روی توپ‌ها نوشته شده و در ظرف‌ها ریخته می‌شوند. استخراج ni نمونه اطلاعات که حاوی تعداد si زیرمجموعه از ویژگی‌هاست مانند انتخاب ni توپ برچسب خورده خواهد بود. ویژگی‌های داخل ظرف ویژگی‌های بالقوه نامیده می‌شوند تا از ویژگی‌های خارج شده متمایز گردند.

بنابراین می‌توان احتمال محاسبه شده را به این صورت تعریف نمود. اگر دو رشته si و sj دارای Pi و Pj ویژگی بالقوه هستند و در مجموعه اطلاعات استخراج شده Di و Dj ظاهر شده‌اند که |Di| = ni و |Dj| = nj و این اطلاعات k ویژگی استخراج شده مشترک دارند. احتمال این که si و sj هم‌مرجع باشند به صورت نشان داده شده در رابطه ۲ محاسبه می‌شود. توجه داشته باشید که احتمال $R_{i,j}^{t}$ بستگی به دو پارامتر Pi و Pj دارد. به دلیل این که ابهام‌زدایی معنایی به صورت بی‌ناظر امکان استفاده از داده‌های برچسب‌خورده به صورت دستی را برای تخمین این پارامترها ندارد، پارامترهای مذکور بر مبنای تعداد دفعاتی که هر کدام از دو رشته si و sj استخراج شده‌اند در نظر گرفته می‌شود.

[20] Extracted Shared Property Model(ESP)

$$P\left(R_{i,j}^{t} \mid D_i, D_j, P_i, P_j\right) = \frac{P(k|n_i, n_j, P_i, P_j, S_{i,j} = P_{min})}{\sum_{k \le S_{i,j} \le P_{min}}^{Si,j} P(k|n_i, n_j, P_i, P_j, S_{i,j})}$$

رابطه2

۲,۱,۴,۱,۲.الگوریتم خوشه‌بندی Resolver

روش‌های بی‌ناظر ابهام‌زدایی معنایی عموماً با مفهوم خوشه‌بندی گره خورده‌اند. در ابهام‌زدایی معنایی به روش بی‌ناظر نیازمند یک الگوریتم خوشه‌بندی می‌باشد تا بتواند با حجم عظیم رشته‌های استخراج شده از وب که به صورت خلوت و با ابعاد و المان‌های بسیار بالا می‌باشد مواجه شود. یک روش استاندارد حریصانه در خوشه‌بندی این است که هر جفت داده را با هم مقایسه کنیم و سپس جفت داده‌های نزدیک به هم را با یکدیگر در یک خوشه ادغام نماییم. این روش حریصانه برای مسئله ما مناسب است زیرا از کوچک‌ترین اندازه خوشه شروع کرده و تا جایی که نیاز باشد به ادغام خوشه‌ها می‌پردازد.

الگوریتم خوشه‌بندی Resolver نیز ویرایشی از همین الگوریتم خوشه‌بندی حریصانه تجمعی می‌باشد. تمایز کلیدی روش استفاده شده که اجازه تعمیم به فضاهای با ابعاد بالا و خلوت و همچنین محموعه عظیم المان‌ها را می‌دهد. الگوریتم استاندارد خوشه‌بندی حریصانه با مقایسه دو به دوی نمونه‌ها شروع می‌کند که همین مرحله اول به ازای N نمونه $O(N^2)$ خواهد بود. روشی که Resolver استفاده می‌نماید این میزان را به O(NLogN) کاهش داده است. الگوریتم ۱ خوشه‌بندی استفاده شده را در این سامانه تشریح می‌نماید.

روش ابتکاری به کار گرفته شده در این الگوریتم این است که فرض نموده‌است که نمونه‌هایی که هیچ ویژگی مشترکی ندارند ارزش مقایسه نخواهند داشت. در نتیجه الگوریتم تنها نمونه‌هایی را با هم مقایسه می‌نماید که ویژگی مشترک با هم داشته باشند. همچنین این فرض نیز در نظر گرفته شده است که نمونه‌هایی که یک ویژگی را با تعداد بسیار زیادی از خوشه‌ها مشترک دارند نیز ارزش مقایسه ندارند، زیرا عموما ویژگی‌هایی که

```
E := { e = (r, a, b) | (r, a, b) is an extracted assertion }
S := { s | s appears as a relation or argument string in E }
Cluster := {}
Elements := {}
1. For each s ∈ S:
        Cluster[s] := new cluster id
        Elements[Cluster[s]] := { s }
2. Scores := {} , Index := {}
3. For each e = (r, a, b) ∈ E:
        property := (a, b)
        Index[property] := Index[property] ∪ { Cluster[r] }
        property := (r, a)
        Index[property] := Index[property] ∪ { Cluster[b] }
        property := (r, b)
        Index[property] := Index[property] ∪ { Cluster[a] }
4. For each property p ∈ Index:
        If | Index[p] | < Max:
                For each pair { c1, c2 } ⊂ Index[p]:
                        Scores[ { c1, c2 } ] := similarity(c1, c2)
5. Repeat until no merges can be performed:
        Sort Scores
        UsedClusters := {}
        Repeat until Scores is empty or top score < Threshold:
                { c1, c2 } := removeTopPair(Scores)
                If neither c1 nor c2 is in UsedClusters:
                        Elements[c1] := Elements[c1] ∪ Elements[c2]
                        For each e ∈ Elements[c2]:
                                Cluster[e] := c1
                delete c2 from Elements
                UsedClusters := UsedClusters ∪ { c1, c2 }
        Repeat steps 2-4 to recalculate Scores
```

الگوریتم ۱: الگوریتم خوشه‌بندی سامانه RESOLVER [۱]

با تعداد بسیار زیاد نمونه‌های دیگر مشترک هستند، بیانگر ارتباط معنایی و هم‌مرجعی بین نمونه‌ها نخواهند بود.

۲،۱،۴،۲. ابهام زدایی با استفاده از خوشه‌بندی رابطه‌ای

در [۳۸] یک روش بی‌ناظر برای ابهام‌زدایی و استخراج شبکه معنایی از متن خام با استفاده از اطلاعات استخراج شده معرفی شده است. این سامانه اطلاعات استفاده شده توسط سامانه‌های استخراج اطلاعات به عنوان ورودی خود دریافت می‌نماید و سپس با خوشه‌بندی همزمان موجودیت‌ها و رابطه‌ها، مفاهیم حقیقی و رابطه‌های بین آن‌ها را ابهام زدایی می‌نماید.

این سامانه بر مبنای منطق مارکوف[۲۱][۳۹] بنا شده است و از مدل خوشه‌بندی رابطه‌ای چندگانه[۲۲][۴۰] استفاده می‌نماید. همچنین این سامانه از اطلاعات استخراج شده توسط سامانه TextRunner به عنوان ورودی استفاده نموده سعی در سازگاری کامل با این سامانه برای کار داشته است. این سازگاری باعث می‌شود که بتوان از جریان اطلاعات بین دو سامانه به خوبی بهره برده و تحلیل داده به صورت مشترک را برای دستیابی به کارآیی و دقت بهتر در نظر داشت[۴۱].

در این سامانه، خوشه‌های مربوط به موجودیت‌ها که در هر مرحله ایجاد شده‌اند برای تولید خوشه‌های مربوط به روابط مورد استفاده قرار می‌گیرند و بالعکس. این روش خوشه‌بندی همزمان نتایج بهتری را از خوشه‌بندی موجودیت‌ها و روابط به صورت جداگانه در بر خواهد داشت.

مدلی که این سامانه از آن استفاده می‌نماید، بر مبنای منطق مرتبه دوم مارکوف بنا شده است که در آن متغیرها می‌توانند روابط و یا موجودیت‌ها باشند. استفاده از منطق مرتبه دوم مارکوف باعث می‌شود که گزاره‌های پایه همراه با تمامی گزاره‌های ممکن و نمادهای ثابت همراه شوند، و این مسئله باعث شده است که مدل فشرده‌تری نسبت به منطق مرتبه اول مارکوف بدست آید.

این سامانه اطلاعات استخراج شده را به صورت $r(x,y)$ در نظر می‌گیرد که در آن موجودیت‌های x و y با یکدیگر رابطه r را دارند. همچنین برای سادگی از نمادهای γ_i و Γ_i به ترتیب برای بیان خوشه و عملیات خوشه‌بندی رشتهi استفاده می‌نماییم.

مسئله یادگیری در این سامانه شامل یافتن $\Gamma = (\Gamma_r,\Gamma_x,\Gamma_y)$ می‌باشد که میزان احتمال شرطی $P(\Gamma|R) \propto P(\Gamma,R) = P(\Gamma)P(R|\Gamma)$ را بیشینه نماید. در این احتمال R بردار صحتی است که به اطلاعات استخراج شده $r(x,y)$ منتسب شده است. در این سامانه برای مولفه شباهت $P(R|\Gamma)$ از یک شبکه منطق مارکوف[۲۳] و برای مولفه احتمال پیشینی $P(\Gamma)$ از یک شبکه منطق مارکوف استفاده شده است[۳۸].

این سامانه از یک روش خوشه‌بندی حریصانه تجمعی پایین به بالا استفاده می‌نماید. الگوریتم ۲ این روش را نشان می‌دهد.

این الگوریتم ابتدا هر نمونه را به عنوان یک خوشه در نظر می‌گیرد و سپس نمونه‌های کاندید را برای ادغام تشخیص داده، و برای هر کدام از این نمونه‌ها میزان تغییر در احتمال شرطی را در صورت ایجاد ادغام محاسبه می‌نماید. اگر ادغام دو خوشه باعث بهبود احتمال شرطی شود، این نمونه‌ها در یک لیست مرتب قرار می‌گیرند. سپس سامانه در لیست مرتب تولید شده حرکت نموده، بهترین ادغام را ابتدا انجام داده و سایر خوشه‌هایی که

[21] Markov Logic
[22] Multiple Relational Clustering (MRC)
[23] Markov Logic Network (MLN)

شامل ادغام اخیر بوده‌اند را نادیده می‌گیرد. به این ترتیب، به صورت افزایشی عمل ادغام خوشه‌ها تا زمانی که ادغام دیگری که منجر به بهبود احتمال شرطی شود ممکن نباشد، ادامه خواهد یافت.

```
function SNE(S r , S x , S y , R)
        inputs: S r , set of relation symbols
                S x , set of object symbols that appear as first arguments
                S y , set of object symbols that appear as second arguments
                R, ground r(x, y) atoms formed from the symbols in S r , S x , and S y
        output: a semantic network, {(γr , γx , γy ) ∈ Γr × Γx × Γy : (γr , γx , γy ) contains
        atleast one true ground atom}
for each i ∈ {r, x, y}
        Γ i ← unitClusters(S i )
mergeOccurred ← true
while mergeOccurred
        mergeOccurred ← false
        for each i ∈ {r, x, y}
                CandidateMerges ← ∅
                for each (γ, γ′) ∈ Γ i × Γ i
                        ΔP ← change in P ({Γr , Γx , Γy }|R) if γ, γ′ are merged
                        if ΔP > 0, CandidateMerges ← CandidateMerges ∪ {(γ, γ′)}
                sort CandidateMerges in descending order of ΔP
                MergedClusters ← ∅
                for each (γ, γ′) ∈ CandidateMerges
                        if γ ∉ MergedClusters and γ′ ∉ MergedClusters
                                Γ i ← (Γ i \ {γ, γ′}) ∪ {γ ∪ γ′}
                                MergedClusters ← MergedClusters ∪ {γ} ∪ {γ′}
                                mergedOccurred ← true
return {( γr , γx , γy) ∈ Γr × Γx × Γy : (γr , γx , γy) contains at least one true ground atom}
```

الگوریتم ۲: الگوریتم خوشه‌بندی رابطه‌ای[۳۸]

در روش فوق، این مشکل وجود دارد که الگوریتم با تولید تمامی جفت‌های ممکن از خوشه‌ها، به ویژه در مراحل ابتدایی، از نظر مرتبه زمانی به شدت روش را کند نماید. برای حل این مشکل، از مفهومی به نام سایبان[۲۴][۴۲] استفاده شده است. در واقع سایبان برای رابطه r به مجموعه‌ای از خوشه‌ها گفته می‌شود که به ازای همه روابط و موجودیت‌های داخل خوشه‌ها، نمونه اطلاعات صحیحی شامل این روابط و موجودیت‌ها وجود داشته باشد. در این الگوریتم خوشه‌هایی به عنوان کاندید برای ادغام قرار می‌گیرند که سایبان آن‌ها بزرگ‌تر از یک پارامتر ثابت CanopyMax نباشد. این پارامتر تعداد جفت‌های خوشه‌ای که برای ادغام کاندید می‌شوند محدود شده و الگوریتم از نظر مرتبه زمانی بهبود یابد. در واقع با استفاده از این ابتکار تنها جفت‌هایی که احتمالا ادغامشان باعث بهبود خواهد شد، کاندید می‌شوند.

[24] Canopy

۲,۱,۴,۳.سامانه DIRT

سامانه DIRT[۲۵]، یک سامانه تشخیص قواعد استلزام از متن خام به روش بی‌ناظر است[۴۳]. به عنوان مثال خروجی این روش به صورت "X is author of Y" نتیجه می‌دهد "X wrote Y" خواهد بود. الگوریتمی که در این سامانه استفاده می‌شود بر مبنای این فرض ااست که کلماتی که در یک زمینه استفاده شده‌اند احتمالا هم‌معنا خواهند بود[۳۴]. البته تفاوت عمده این سامانه با سایر روش‌هایی که از این فرض استفاده می‌کنند در این است که این سامانه به جای استفاده از کلمات به عنوان زمینه، از درخت وابستگی جملات به عنوان زمینه استفاده نموده و در واقع مسیرهای این درخت را در روابط متفاوت با هم مقایسه می‌نماید.

تشابه بین دو مسیر در درخت وابستگی به این صورت محاسبه می‌شود که دو مسیر با هم مشابه هستند، اگر تعداد زیادی از خصیصه‌های مشترک داشته باشند. البته تمامی خصیصه‌ها در این مجموعه ارزش یکسانی نخواهند داشت. برای مثال کلمه he بسیار رایج‌تر از کلمه‌ای مانند Adam است. اما دو مسیر که خصیصه (SlotX, he) را اشتراک دارند بسیار کم‌ارزش‌تر از اشتراک خصیصه (SlotX, Adam) است.

این سامانه برای محاسبه معیار شباهت از معیار اطلاعات متقابل[۲۶] بین خصیصه و مسیر استفاده نموده است. برای حل چالش مرتبه زمانی مقایسه تمامی مسیرهای ممکن با هم برای یافتن مشابه‌ترین مسیرها، این سامانه از یک روش سه مرحله‌ای ابتکاری استفاده نموده است. برای هر مسیر، در مرحله اول تمامی مسیرهایی که حداقل یک خصیصه مشترک با این مسیر دارند، بازیابی می‌شوند. در مرحله دوم از مجموعه مسیرهای کاندید انتخاب شده، مسیرهایی که تعداد خصیصه‌های مشترک آن‌ها با مسیر مورد نظر کم‌تر از میزان ثابت باشند حذف می‌شوند و در مرحله سوم تنها میزان تشابه مجموعه باقیمانده با مسیر مورد نظر محاسبه می‌شود.

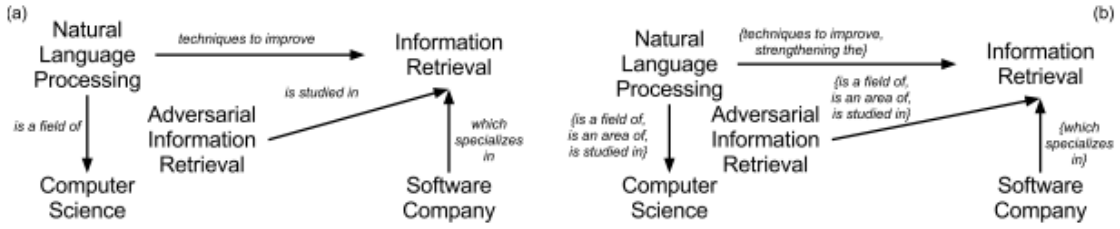
۲,۱,۴,۴. سامانه WiseNet

سامانه WiseNet با به خدمت گرفتن روش‌های استخراج آزاد اطلاعات و روش‌های بهره‌برداری از دانش[۲۷]، یک شبکه معنایی بر مبنای ویکی‌پدیا را تولید می‌نماید[۴۴]. این الگوریتم روابط را از صفحات ویکی‌پدیا استخراج می‌نماید و این روش‌ها را در مجموعه‌هایی از عبارات رابطه‌ای هم‌معنا دسته‌بندی می‌نماید. سپس کلاس‌های معنایی را به آرگومان‌های روابط نسبت می‌دهد. خروجی این سامانه یک شبکه معنایی مبتنی بر صفحات ویکی‌پدیا خواهد بود که برچسب‌گذاری شده و ابهام‌زدایی شده می‌باشد.

برای خوشه‌بندی روابط هم‌معنا در این سامانه، هر رابطه به صورت دو بردار تعریف می‌شود که مولفه‌های تعداد ظاهر شدن پرتکرارترین کلماتی که سمت چپ و سمت راست این روابط ظاهر شده‌اند در یک پیکره متنی بزرگ خواهد بود. سپس معیار شباهت برای مقایسه روابط و خوشه‌بندی، میانگین هارمونیک بین معیار تشابه

[25] Discovery of Inference Rules from Text
[26] Mutual Information
[27] Knowledge Aqcuisition

کوسینوسی بین این دو بردار خواهد بود. خروجی این مرحله از سامانه، مجموعه‌هایی از روابط هم‌معنا خواهد بود. در شکل ۱ شبکه معنایی حاصل را بعد از مرحله اول و مرحله دوم مشاهده می‌نمایید.



شکل2 : (a) مثالی از خروجی وایزنت بعد از مرحله اول (b) مثالی از خروجی وایزنت بعد از مرحله دوم[۴۴]

در مرحله سوم، موجودیت‌ها در کلاس‌های معنایی که از ویکی‌پدیا گرفته شده‌است قرار داده می‌شوند. این سامانه کلاس‌های معنایی را به وسیله موضوعات ویکی‌پدیا مدل می‌کند. به این ترتیب که به ازای هر آرگومان، با استفاده از جستجوی عمقی[۲۸] با عمق مشخص موضوعات مربوط به این آرگومان در سلسله مراتب موضوعات بازدید شده و تعداد دفعاتی که هر موضوع بازدید شده، شمرده می‌شود. این تعداد دفعات می‌تواند امتیازی برای رتبه‌بندی موضوعات به عنوان کلاس معنایی مجموعه روابط باشند.

در جدول ۲ نمونه‌ای از اطلاعات حاصل از این سامانه به طور کامل مشاهده می‌شود:

جدول2 : مثالی از اطلاعات حاصل از سامانه وایزنت شامل کلاس‌ها معنایی موجودیت‌ها و روابط[۴۴]

Left Semantic Classes	Relation Synset	Right Semantic Classes
Scientific disciplines, Applied sciences, Academic disciplines	is a field of, is an area of, is studied in	Scientific disciplines, Applied Science, . . ., Academic disciplines
Peaple, Academics, Students, …, Education and training occupations	have a BSc in, hold a B.Sc. degree in, possess an undergraduate degree in	Academic disciplines, Science , ,. . . Scientific disciplines
Peaple, Society, …, Dictional organizations	assist the, aid the, help the	People, . . ., Society

۲,۲. ارزیابی اطلاعات

ارزیابی اطلاعات استخراج شده به روش بی‌ناظر به طور ذاتی یک مسئله چالش‌برانگیز است. زمانی که روش‌های دست‌ساز، با ناظر و یا نیمه-نظارتی را برای استخراج اطلاعات استفاده می‌نمایم، در واقع از قبل مجموعه

[28] Depth-first Search (DFS)

موجودیت‌ها و روابطی که می‌خواهیم استخراج کنیم را مشخص نموده‌ایم؛ بنابراین اگر چیز متفاوتی استخراج شود، آن را به عنوان خطا قلمداد خواهیم نمود.

اما زمانی که صحبت از استخراج بی‌ناظر اطلاعات، نظیر آن‌چه در استخراج آزاد اطلاعات با آن مواجه هستیم باشد، به این ترتیب ما باید از خود داده بپرسیم که کدام کلاس‌های استخراج شده خوب هستند! به این ترتیب، به سادگی نمی‌توانیم در مورد صحت و یا عدم صحت این اطلاعات قضاوت کنیم. همچنین اگر در مورد یک نمونه از اطلاعات، با کلاس‌های متفاوتی مواجه شویم، به راحتی نمی‌توان گفت که کدام‌یک از آن کلاس ها درست هستند.

به این ترتیب در استخراج آزاد اطلاعات، یا به طور کلی در استخراج اطلاعات در غیاب مداخله انسانی، ما به عنوان خروجی با انبوهی از اطلاعات نامطمئن مواجه خواهیم بود. در ادامه چند روش را برای مشخص کردن میزان صحت اطلاعات استخراج شده، و به عبارت دیگر معیارهایی را برای ارزیابی و محاسبه احتمال صحت اطلاعات استخراج شده معرفی می‌کنیم.

۲،۲،۱. ارزیابی بر مبنای فرکانس

در نگاه اول به نظر می‌رسد که هرچه یک اطلاع بیشتر استخراج شده باشد، احتمال صحت آن نیز بیشتر است. برای استفاده از این منطق ساده در تبدیل میزان تکرار استخراج به احتمال صحت اطلاع، رابطه $1 - 0.5^c$ مورد استفاده قرار گرفته است. در این رابطه با افزایش مقدار c که همان میزان تکرار استخراج است، احتمال صحت به شکل لگاریتمی بالا می‌رود[45].

مشکل اصلی این روش ساده، عدم جلوگیری از خطاهای سیستمی است. در واقع این روش هیچ نوع ارزیابی روی روش استخراج اطلاعات ندارد. به این ترتیب وقتی فرایند استخراج به هر دلیل، اشتباهی را تکرار کند، اطلاع استخراج‌شده‌ی حاصل، احتمال بالایی پیدا می‌کند و به هیچ شکلی مورد بررسی مجدد قرار نمی‌گیرد. این مشکل به قدری جدی است که حتی در مواردی، بخش محدودی از پرتکرارترین استخراج‌ها از خروجی حذف می‌شوند تا دقت افزایش یابد.

۲،۲،۲. ارزیابی با جستجو در وب

از موتورهای جستجو هم می‌توان برای محاسبه‌ی احتمال صحت یک اطلاع کمک گرفت. در هنگام توضیح سامانه‌ی KnowItAll گفتیم که این سامانه اطلاعات بدست آمده را در شکل وقوع مورد انتظار جایگزین می‌کند و از موتور جستجو میزان تکرار آن‌ها را می‌پرسد. خروجی موتور جستجو، به میزان باهم‌آیی تبدیل می‌شود و در محاسبه احتمال صحت مورد استفاده قرار می‌گیرد[۴۶].

۲،۲،۳. مدل تکرار

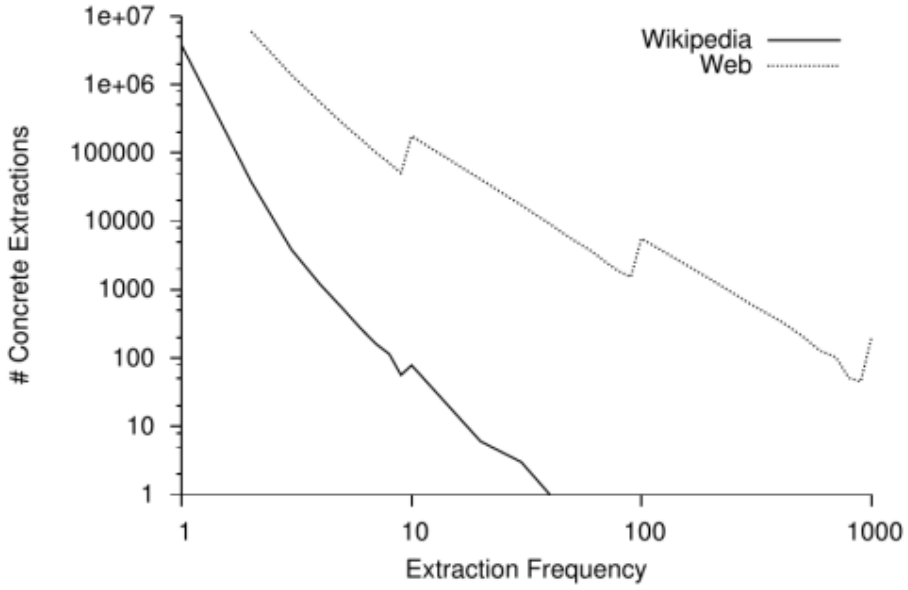
همانطور که گفته شد، می‌توانیم روی استخراج‌های مختلف یک اطلاع در وب حساب کنیم؛ چون با انبوه متون مواجه هستیم. تلاش‌هایی برای حل مساله بدست آوردن احتمال صحت از روی میزان تکرار، با مدل‌سازی

دقیق احتمالاتی شده است[۳۷]. برای این منظور، مساله‌ی استخراج اطلاعات بر مدل توپ‌ها و گلدان‌ها انطباق داده شده و مدلی برای بدست آوردن این احتمال پیشنهاد شده است. برای برقراری این انطباق، هر رابطه‌ای که هدف استخراج است به یک گلدان تشبیه شده و هر اطلاع مورد استخراج با توپی برچسب‌دار نمایش داده شده است. به این ترتیب ما تعداد زیادی گلدان (رابطه) داریم و هر بار که اطلاعی استخراج می‌کنیم، در واقع یک توپ (با جایگزینی) از گلدان‌ها بیرون کشیده‌ایم. پس از پایان استخراج همه‌ی مشاهده‌ی ما برای یک اطلاع استخراج‌شده با برچسب مشخص در این جمله خلاصه می‌شود: «برچسب x بعد از n بار بیرون کشیدن توپ، به تعداد k بار مشاهده شده است». اگر فرض کنیم C تعداد برچسب‌های مورد نظر باشد و E تعداد برچسب‌های خطا را نمایش دهد. در صورتی که تعداد همه‌ی توپ‌های درون گلدان برای r تکرار s محاسبه شود، آنگاه احتمال صحت برچسب (اطلاع استخراج‌شده) مورد نظر از رابطه ۳ محاسبه می‌شود.

$$P(x \in C|O) = \frac{\sum_{r \in num(C)} \left(\frac{r}{s}\right)^k \left(1 - \frac{r}{s}\right)^{n-k}}{\sum_{r' \in num(C \cup E)} \left(\frac{r'}{s}\right)^k \left(1 - \frac{r'}{s}\right)^{n-k}}$$

رابطه3

البته باید توجه داشت که میزان تکرار روابط در متون، با توجه به نوع آن‌ها ممکن است متفاوت باشد. برای نمونه در آزمایشی نشان داده شده است که هر رابطه‌ای که از متن ویکی‌پدیا یک بار استخراج شده، به طور متوسط از پیکره‌ی مربوط به وب، تقریبا چهار بار بیرون آمده است[۷]. همچنین در شکل ۳ مشاهده می‌کنیم که در متن ویکی‌پدیا تعداد وقوع روابط بسیار کمتر و البته با همین تکرار کم بسیار قابل اطمینان است.



شکل3: شناسایی میزان تکرار روابط در وب و متن و ویکی‌پدیا[۷]

فصل سوم:

کاربردها

۳.کاربردها

تا به این جا به بررسی اجمالی روش‌های استخراج اطلاعات پرداخته‌ایم، سپس دو چالش اساسی در استخراج با کیفیت اطلاعات، یعنی ابهام و ارزیابی را معرفی نموده و روش‌های مواجه با این چالش‌ها را بررسی نمودیم. بنابراین در این مرحله با مجموعه‌ای از اطلاعات، یعنی روابط میان موجودیت‌ها، که هر رابطه در واقع گروهی از روابط هم‌معنا می‌باشد، و هر موجودیت هم در واقع گروهی از نام‌های متفاوت برای یک موجودیت را در شامل می‌شود. همچنین اطلاعات با احتمال صحت کم‌تر و در واقع اطلاعات کم‌ارزش‌تر هم از این مجموعه اطلاعات حذف شده‌اند.

در واقع به عنوان خروجی این سامانه، ما تا حدی زیادی به آن‌چه که به عنوان استخراج دانش از متن زبان طبیعی نامیده می‌شود، نزدیک شده‌ایم. در این بخش به بررسی زمینه‌هایی که از خروجی این روش‌ها، یعنی اطلاعات با حجم عظیم و صحت قابل قبول، می‌توانیم در آن‌ها استفاده کنیم، خواهیم پرداخت.

۳,۱. استخراج دانش

تا به این‌جا بحث ما عموما پیرامون اطلاعات و استخراج آن از متن زبان طبیعی بود. در واقع هدف ما مشخص کردن روابط و موجودیت‌های آرگومان این روابط بود. اما وقتی صحبت از استخراج دانش به میان می‌آید، تمایل داریم که سطح عمیق‌تری از اطلاعات را داشته باشیم. اولین گام برای رسیدن به فهم از روی دانسته‌های سطحی حاصل از استخراج اطلاعات، شناسایی موجودیت‌ها و نوع آن‌ها و همچنین شناسایی و گروه‌بندی روابط هم‌معنی خواهد بود.

یکی از روش‌های استخراج دانش، استفاده از منابع دانش خارجی نظیر ویکی‌پدیا برای این کار می‌باشد.[۴۷] به بررسی روش‌های استخراج دانش از منابع متنی وب نظیر ویکی‌پدیا پرداخته است. به عنوان اولین مرحله برای استخراج دانش از متن، نیاز داریم تا موجودیت‌های متن را استخراج نموده و نام‌های هم‌معنا که به یک موجودیت مربوط می‌شوند را با هم در یک گروه قرار دهیم. البته هر نام ممکن است در چند گروه قرار گیرد. به عنوان مثال نام «حسن روحانی» می‌تواند در گروه‌هایی متفاوت همراه با «رئیس جمهور»، «سیاست‌مدار» و «استاد دانشگاه» قرار گیرد. این کار بسیار مشابه با مسئله «یادگیری هستان‌شناسی» خواهد بود[۴۷]. روش‌هایی وجود دارند که برای این کار از منابع دانش خارجی نظیر وردنت و یا ویکی‌پدیا استفاده می‌کنند. که به عنوان معروف‌ترین آن‌ها می‌توان از Yago و WikiTaxonomy نام برد[۳], [۴۸], [۴۹]. Yago با استفاده از وردنت و همچنین دسته‌های موضوعی ویکی‌پدیا روشی را برای تولید پایگاه دانش به طور خودکار ارائه نموده است.

از دیگر نمونه‌های این کاربرد می‌توان به سامانه SOFIE اشاره کرد که روشی یکپارچه برای استخراج دانش از متن بر مبنای استخراج اطلاعات به صورت خودسامان‌ده[۲۹] ارائه نموده است[۵۰]. البته SOFIE نیاز به یک

[29] Self-Organizing

هستان‌شناسی موجود به عنوان هسته‌ی اولیه خواهد داشت و هدف این روش توسعه و گسترش دانش هستان‌شناسی‌های موجود می‌باشد.

سامانه‌ی ALICE، که در فصل قبل به شرح روش آن پرداخته‌ایم، یک عامل یادگیری مادام‌العمر می‌باشد که هدف آن کشف و استخراج خودکار دانش، یعنی مجموعه‌ای از مفاهیم و واقعیت‌ها، و همچنین تعمیم آن‌ها، به طور مستقیم از متن زبان طبیعی موجود در وب می‌باشد[۱۳]. این سامانه برای روش خود نیاز به پیکره‌ای با دامنه‌ی موضوعی خاص، یک منبع دانش زمینه می‌باشد. همچنین ALICE از سامانه استخراج اطلاعات TextRunner به عنوان موتور استخراج اطلاعات بهره می‌برد.

۳,۲. استنتاج از متن زبان طبیعی

استنتاج از متن زبان طبیعی[۳۰] مسئله تشخیص این است که آیا یک فرضیه که به زبان طبیعی بیان شده است، می‌تواند به طور منطقی از فرضیه‌های موجودی که آن‌ها نیز به زبان طبیعی بیان شده‌اند، استخراج شود. استنتاج یکی از مباحث اصلی در هوش مصنوعی می‌باشد که در پنج دهه اخیر سپری از عمر این رشته، پژوهش‌گران پیشرفت‌های عظیمی را در توسعه روش‌های خودکار استنتاج منطقی به وجود آورده‌اند. اما چالش استنتاج از زبان طبیعی، همچنان به عنوان یک چالش پیش رو، مسئله‌ای کاملا متفاوت از آن‌چه که در این زمینه پیشرفت شده است می‌باشد. زیرا این مسئله علاوه بر روش‌های استنتاج منطقی، نیازمند یک منبع دانش معنایی واژگانی و همچنین مدیریت مسئله تنوع در عبارات زبان طبیعی می‌باشد.

بنابراین یکی از الزامات توسعه در استنتاج زبان طبیعی، این است که بتوانیم دانش را از متن زبان طبیعی استخراج کنیم. در واقع در این‌جا ما نیاز داریم همه گزاره‌های موجود در متن را بفهمیم و بتوانیم با فهم‌های بدست آمده از متن، استنتاج انجام دهیم.

سامانه Sherlock روشی را ارائه نموده است تا از انبوه اطلاعات استخراج شده توسط سامانه‌های استخراج اطلاعات، برای تولید دانش مورد نیاز استنتاج استفاده شود[۵۱]. این سامانه وظیفه استخراج گزاره‌های شرطی را بر عهده دارد. در واقع هدف این سامانه تبدیل اطلاعات استخراج شده به گزاره‌های منطقی به فرم هورن[۳۱] است. سامانه‌ی دیگری با نام Holmes نیز براس استنتاج از این گزاره‌های شرطی توسعه داده شده است[۵۲].

اولین گام برای رسیدن به دانش از روی اطلاعات. شناسایی موجودیت‌ها و نوع آن‌هاست. روش‌های حل این مسئله را در فصل گذشته با عنوان روش های ابهام‌زدایی اطلاعات استخراج شده بررسی نمودیم. اما سامانه Sherlock از ساده‌ترین روش ممکن برای این کار، یعنی استفاده از الگوهای متنی و تعریف قواعدی بر این مبنا استفاده نموده است. این روش کاملا مبتنی بر قاعده[۳۲] می‌باشد و می‌توان آن را ابتدایی‌ترین روش انجام این

[30] Natural Language Inference (NLI)
[31] Horn clause
[32] Rule-base

کار دانست. برای مثال الگوی «**نوع** از قبیل **نمونه**» می‌تواند یک الگو باشد که نمونه‌هایی نظیر «**شهرهایی** از قبیل تهران و اصفهان» منجر به شناسایی «تهران» و «اصفهان» به عنوان نمونه‌هایی از نوع شهر می‌شود.

پس از این مرحله، این سامانه اقدام به کشف گزاره‌های مربوط به این وجودیت‌ها می‌نماید. همان‌طور که ذکر شد، Sherlock با گزاره‌های به فرم هورن کار می‌کند و سعی می‌کند آن‌ها را یاد بگیرد. این سامانه با تولید همه گزاره‌های ممکن برای نوع‌های پرتکرار، سعی می‌کند گزینه‌های قابل بررسی را تولید کرده و سپس صحت آن‌ها را از روی داده‌ها تحقیق کند. ابتدا به طور پیش‌فرض همه‌ی گزاره‌های تولید شده غلط محسوب می‌شوند. سپس برای بررسی صحبت یک گزاره، از مفهوم ارتباط آماری[۳۳] استفاده می‌شود[۵۳].

در شکل ۴ نمونه‌هایی از گزاره‌های استخراج شده توسط این سامانه مشاهده می‌شود:

IsHeadquarteredIn(Company, State) :- IsBasedIn(Company, City) ∧ IsLocatedIn(City, State);

Contains(Food, Chemical) :- IsMadeFrom(Food, Ingredient) ∧ Contains(Ingredient, Chemical);

Reduce(Medication, Factor) :- KnownGenericallyAs(Medication, Drug) ∧ Reduce(Drug, Factor);

ReturnTo(Writer, Place) :- BornIn(Writer, City) ∧ CapitalOf(City, Place);

Make(Company1, Device) :- Buy(Company1, Company2) ∧ Make(Company2, Device);

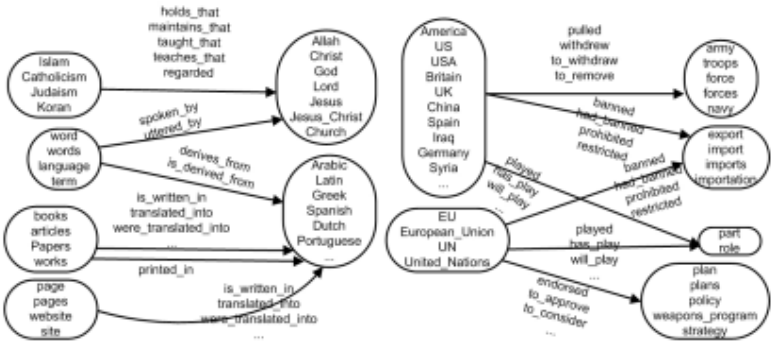
شکل ۴: نمونه‌ای از گزاره‌های استخراج شده از سامانه SHERLOCK

۳,۳. استخراج خودکار شبکه معنایی

یکی از اهداف مورد نظر برای هوش مصنوعی، ساختن عامل‌های خودمختاری است که بتوانند فهمی از متن داشته زبان طبیعی داشته باشند. یکی از تلاش‌هایی که در این زمینه در پردازش زبان طبیعی توسط محققین بسیاری دنبال شده است، ساخت سامانه‌هایی است که متن زبان طبیعی را به به شبکه‌ای از مفاهیم تجزیه نماید و منبع دانشی را بر مبنای این شبکه از مفاهیم و روابط میان آن‌ها تولید کند.

با توجه به منبع عظیمی از دانش که در به صورت متن زبان طبیعی در وب موجود است، ارائه روشی برای استخراج خودکار این شبکه معنایی از متن وب می‌تواند بسیار مطلوب باشد. در [۳۸] روشی برای استخراج خودکار شبکه معنایی از متن زبان طبیعی به صورت بی‌ناظر، با استفاده از خوشه‌بندی رابطه‌ای ارائه شده است که در فصل گذشته پیرامون آن به بحث پرداختیم. این روش اطلاعات استخراج شده توسط سامانه استخراج آزاد اطلاعات TextRunner را به عنوان ورودی دریافت نموده و شبکه‌ی معنایی از موجودیت‌ها و گروه‌های روابط میان آن‌ها، نظیر آن‌چه در شکل ۵ مشاهده می‌کنید، تولید می‌نماید.

[33] Statistical Relevance



شکل5 : شبکه معنایی حاصل از اطلاعات استخراج شده به وسیله سامانه Textrunner [۳۸]

سامانه WiseNet [۴۴] تلاش دیگری در استخراج شبکه معنایی از ویکی‌پدیاست. این سامانه با به خدمت گرفتن روش‌های استخراج آزاد اطلاعات و روش‌های بهره‌برداری از دانش[۳۴]، یک شبکه معنایی بر مبنای ویکی‌پدیا را تولید می‌نماید. این سامانه روابط را از صفحات ویکی‌پدیا استخراج می‌نماید و این روش‌ها را در مجموعه‌هایی از عبارات رابطه‌ای هم‌معنا دسته‌بندی می‌نماید. سپس کلاس‌های معنایی را به آرگومان‌های روابط نسبت می‌دهد. خروجی این سامانه یک شبکه معنایی مبتنی بر صفحات ویکی‌پدیا خواهد بود که برچسب‌گذاری شده و ابهام‌زدایی شده می‌باشد.

۳,۴. پاسخ به پرسش‌ها

یکی از کاربردهای اصلی استخراج اطلاعات، فراهم کردن پاسخ در سامانه‌هایی است که به کاربر اجازه‌ی پرسش می‌دهند. همانطور که در توضیح سامانه TextRunner اشاره کردیم، انجام یک بازیابی اطلاعات ساده روی روابط حاصل از فرایند استخراج تا حد زیادی کاربر را به پاسخ سوال خود نزدیک می‌کند. نیازی که باعث می‌شود چنین جستجویی ارزشمند شود، محدود بودن موتورهای جستجوی وب در پاسخ به سوالات است. اگرچه نحوه دریافت پرس‌وجو از کاربر در موتورهای جستجوی کنونی بسیار آسان و مناسب کاربر است، کاربر نمی‌تواند پاسخ پرسش‌های زیر را از این موتورها دریافت کند:

- نحوه رابطه میان دو موجودیت (مثل ابن‌سینا و الکل)
- پاسخ کوتاه یک پرسش روشن (مثل تاریخ تولد یک دانشمند)
- فهرستی از موجودیت‌هایی با ویژگی مشترک (مثل اسامی شهرهای یک کشور)

این پرسش‌ها نمونه‌ای از سوال واقعی کاربر هستند که برای پرسش آن‌ها از موتورهای جستجو باید از کلید واژه‌های مربوط به هر کدام استفاده کند و جواب را خود از میان نتایج بدست آمده بیرون بکشد.

[34] Knowledge Aqcuisition

در [۵۴] سامانه‌ای برای پاسخ به پرسش‌ها بر مبنای اطلاعات استخراج شده معرفی شده است. در این سامانه پاسخ به پرسش‌ها به عنوان یک مسئله یادگیری ماشین در نظر گرفته شده و الگوریتمی برای پاسخ به سوالات به صورت مستقل از دامنه موضوعی ارائه نموده است که سوالات را به جستارهایی[۳۵] روی پایگاه داده اطلاعات استخراج شده از وب تبدیل می‌نماید. این سامانه با استفاده از یک پیکره بزرگ از پرسش نوشته شده توسط کاربران، نشان داده است که می‌تواند یک واژه‌نامه معنایی و همچنین یک تابع خطی را برای پاسخ به سوالات با استفاده از اطلاعات استخراج شده، بدون هیچ‌گونه نمونه سوال دست‌ساز تولید نماید.

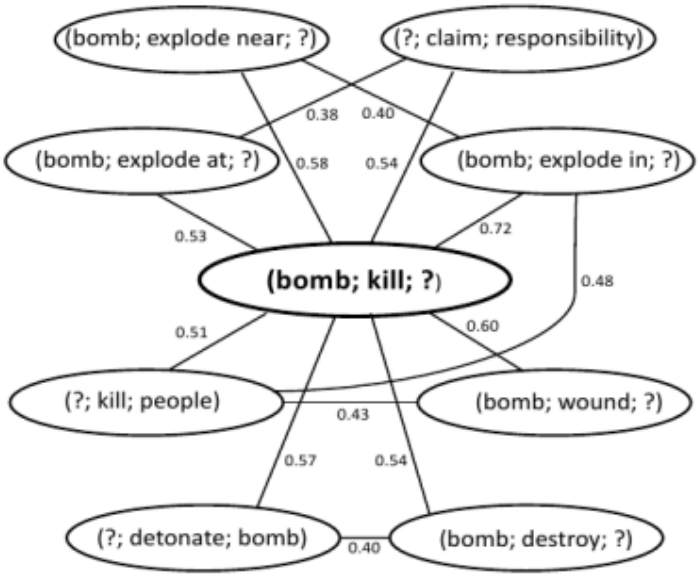
۳,۵. مدل زبانی رابطه‌ای

استخراج آزاد اطلاعات را می‌توان به معنی تبدیل متن خام به مجموعه‌ای از روابط دانست. در این نگاه، متن دیگر دنباله‌ای از واژه‌ها نیست و ما با دنباله‌ای از روابط مواجه هستیم. مدل زبانی رابطه‌ای برای کمی کردن ترتیب قرارگیری این روابط ارائه شد[۵۵] .تصور ما از مدل زبانی[۳۶] مشاهده‌ی ساده‌ای بود که وقتی یک واژه را می‌بینیم، واژه‌ی بعدی آن را می‌توانیم تا حدی حدس بزنیم؛ یعنی مشاهده‌ی یک واژه، روی احتمال بروز واژه‌های دیگر تاثیر می‌گزارد. مدل زبانی رابطه‌ای هم به همین معنی می‌پردازد و فقط تفاوت در استفاده از رابطه‌ها به جای واژه‌هاست.

اگر احتمال بروز یک رابطه را به شرط مشاهده‌ی دیگر روابط پیش از آن محاسبه کنیم، مدل مورد نظر ایجاد می‌شود. به نظر می‌رسد در این مدل، اطلاعاتی در سطح معنای متن نهفته است و می‌توان از آن در شباهت‌یابی میان متون بهره برد. با استفاده از این مدل می‌توان احتمال وقوع یک متن را بررسی کرد؛ چنان که پیشنهاد دهندگان این مدل، احتمال وقوع متون خبری واقعی را محاسبه کرده‌اند. مشاهدات ایشان نشان می‌دهد که احتمال وقوع یک متن خبری بسیار بیشتر از متنی است که از کنار هم گذاشتن اتفاقی جمله‌های بریده شده از متن، ایجاد می‌شود. ایده‌ی دیگر این پژوهشگران، مطرح کردن گراف روابط است که یال‌های آن میزان باهم‌آیی روابط را نشان می‌دهد، یک نمونه از این گراف را در شکل ۶ می‌بینیم.

در این شکل مشخص است که مثلا پس از مطرح کردن وقوع انفجار بمب در محل، میزان تلفات آن گزارش می‌شود.

[35] Query
[36] Language model



شکل6 : نمایش میزان با هم‌آیی روابط استخراج شده در قالب گراف [۵۵]

۳,۶. شباهت‌یابی معنایی متون

اشاره کردیم که حاصل استخراج آزاد اطلاعات، تبدیل متن به مجموعه‌ای از روابط است. با این نگاه می‌توان روابط میان متون را نیز بازتعریف کرد. یعنی وقتی یک رابطه در دو متن و با آرگومان‌های یکسان تکرار می‌شود، نشانه‌ی بسیار محکمی برای شباهت آن‌هاست. در واقع این دلیل تشابه، بسیار محکم‌تر از بروز همه‌ی واژه‌های یک رابطه در هر دو متن است. نکته‌ی قابل توجه آن است که حتی اگر یک رابطه در دو متن با یک آرگومان مشابه ظاهر شود، باز هم می‌تواند دلیل خوبی برای وجود تشابه میان دو متن باشد.

چالش اصلی نزدیک شدن به این مساله، تنک بودن فضای ویژگی‌های متن است و این مشکل قابل حل نیست مگر با روش‌هایی که تحت عنوان شناسایی روابط و موجودیت‌های مطرح کردیم. امید است که با استفاده از این روش بتوان توصیف جدیدی از متن داشت که قابل استفاده در شباهت‌یابی، رده‌بندی، کشف تقلب و ... است.

فصل چهارم:

خلاصه و جمع‌بندی

۴.خلاصه و جمع‌بندی

در این پژوهش به بررسی روش‌های ابهام‌زدایی و ارزیابی اطلاعات استخراج شده از متن زبان طبیعی پرداختیم. ابتدا مروری اجمالی بر روش‌های استخراج اطلاعات و معرفی آن‌ها پرداختیم. سپس به بررسی چالش‌های پیش روی استخراج اطلاعات، ما نگاه ویژه به استخراج آزاد اطلاعات پرداخته و دو چالش اساسی ابهام در اطلاعات و روش ارزیابی بی‌ناظر میزان صحت اطلاعات استخراج شده را معرفی کرده و آن‌ها را به عنوان مسائل اساسی برای استخراج با کیفیت اطلاعات بررسی کردیم.

ابهام در اطلاعات استخراج شده را به این صورت تعریف نمودیم که در استخراج انبوه از اطلاعات، بسیاری از موجودیت‌ها با نام‌های متفاوتی ظاهر می‌شوند و همچنین روابط نیز ممکن است صورت‌های متفاوتی را در اطلاعات مختلف داشته باشند. بنابراین اغلب در مجموعه اطلاعات استخراج شده، نمونه‌های با اطلاعات یکسان زیادی وجود دارند که صورت‌های متفاوتی در نام موجودیت‌ها، نام روابط و یا هر دو دارند. چالشی که در این حوزه با آن مواجه خواهیم بود، این اطلاعات یکسان را تشخیص داده و در واقع به ازای هر موجودیت، دسته‌ای از نام‌هایی که موجودیت مورد نظر در مجموعه اطلاعات با آن ظاهر شده‌اند، داشته باشیم. و همچنین به ازای هر کدام از روابط، دسته‌ای از روابط هم‌معنا را ایجاد نماییم. با توجه به نگاه ویژه‌ای که به استخراج آزاد اطلاعات از وب داشتیم، گفتیم که روش‌هایی مطلوب ما خواهند بود که به صورت ذاتی بدون مداخله انسان باشند تا امکان گسترش آن‌ها در فضای عظیمی مثل وب وجود داشته باشد. همچنین این روش‌ها باید تا حد ممکن از منابع دانش خارجی استفاده نکنند، زیرا استفاده از منابع دانش خارجی باعث محدود شدن سامانه به دامنه موضوعی منبع دانش خواهد گشت، در حالی که مطلوب ما این است که روشی مستقل از دامنه‌های موضوعی داشته باشیم تا بتوانیم در پیکره‌های عظیم و ناهمگون متنی نظیر وب نیز از این روش استفاده کنیم. همچنین ذکر شد که برخلاف بسیاری از روش‌ها، برای استخراج انبوه اطلاعات مطلوب این است که ابهام‌زدایی بدون استفاده از متن زمینه و تنها با استفاده از اطلاعات موجود صورت گیرد.

روش‌های ابهام‌زدایی معرفی شده در این پژوهش به سه بخش تقسیم شدند. روش‌های ابهام‌زدایی روی موجودیت‌ها، روش‌های ابهام‌زدایی روابط، و روش‌هایی که ابهام‌زدایی را برای هر دو جزء اطلاعات، یعنی روابط و موجودیت‌ها انجام می‌دهند.

سپس به بررسی روش‌های ارزیابی اطلاعات استخراج شده پرداختیم. ذکر کردیم که ارزیابی را می‌توان از دو جنبه بررسی نمود. یکی ارزیابی میزان دقت روش استخراج اطلاعات و دیگری ارزیابی صحت و ارزشمندی اطلاعات استخراج شده و گفتیم که با توجه به ذات بی‌ناظر بودن استخراج آزاد اطلاعات، ارزیابی اطلاعات در غیاب دخالت انسانی و داده‌های دست‌ساز یک چالش قلمداد می‌شود. سپس به بررسی روش‌های ارزیابی اطلاعات استخراج شده پرداختیم. ای روش‌ها اغلب به ارزیابی اطلاعات با جستجو در وب و یا با استفاده از ویژگی انبوه و ناهمگون بودن اطلاعات استخراج شده از پیکره عظیم وب پرداخته بودند.

در بخش بعدی به بررسی کاربردهای استخراج آزاد اطلاعات، با فرض افزایش کیفیت اطلاعات استخراج شده با استفاده از روش‌های ابهام‌زدایی و ارزیابی اطلاعات پرداختیم. در واقع با فرض این که خروجی سامانه با

استفاده از این روش‌ها به صورت مجموعه‌ای از اطلاعاتی می‌باشد که هر جزء آن گروه‌های موجودیت‌ها و یا روابط هم‌معنا خواهند بود و اطلاعات با میزان احتمال صحت پایین نیز در بین آن‌ها حذف شده‌اند؛ سپس با این فرض به معرفی مسائل و کاربردهایی پرداختیم که از این مجموعه می‌توان در حل آن‌ها استفاده نمود. از جمله این مسائل به بررسی استخراج دانش، استنتاج از زبان طبیعی، استخراج خودکار شبکه معنایی، پاسخ به پرسش‌ها، مدل زبانی رابطه‌ای و شباهت‌یابی معنایی متون را بررسی کرده و روش‌های توسعه داده شده بر مبنای استخراج آزاد اطلاعات را در این مسائل معرفی نمودیم.

۵. پیشنهاد

در این پژوهش‌نامه ابتدا روش‌های مختلف استخراج اطلاعات، از جمله روش‌های استخراج هدفمند و آزاد اطلاعات، را به طور اجمالی بررسی نمودیم. سپس تعریفی از اطلاعاتی که مد نظر ماست، به صورت سه‌تایی رابطه و دو آرگومانش تعریف نمودیم. بعد از ذکر کردیم که برای داشتن یک استخراج آزاد با کیفیت، با چالش‌هایی روبرو خواهیم بود که باید راه حلی برای آن‌ها داشته باشیم. ابهام در اطلاعات استخراج شده و همچنین چگونگی ارزیابی میزان صحت اطلاعات را به عنوان دو چالش اساسی معرفی کرده و سپس به راه حل‌های ممکن برای مواجهه با این دو چالش پرداختیم.

با توجه به این که رویکرد اصلی ما از استخراج اطلاعات در این پژوهش، استخراج اطلاعات در سطح وسیع، که آن را استخراج آزاد اطلاعات می‌نامیم، بوده است و روش‌های ارائه شده را نیز با این نگاه ویژه بررسی نمودیم؛ ذکر این نکته لازم می‌باشد که روش‌های ارائه شده برای استخراج آزاد اطلاعات، با توجه به ویژگی‌های خاصی که باید داشته باشند، از جمله ماهیت بی‌ناظر بودن آن‌ها و همچنین عدم استفاده از دانش خارجی تا جای ممکن، نیاز این روش‌ها را به استفاده از خصوصیت‌های زبانی دوچندان می‌نماید. در واقع روش‌هایی که در این حوزه ارائه شده‌اند، سعی می‌نمایند تا حد ممکن مستقل از دامنه موضوعی و روابط باشند، اما این مستقل بودن باعث می‌شود که در بسیاری از روش‌های مورد استفاده، مسئله استقلال از زبان از بین رفته و با استفاده از خصوصیت‌های زبانی، این روش‌ها در برخی مراحل خود کاملا وابسته به زبان عمل می‌نمایند.

با توجه به محدودیت منابع پردازشی زبان طبیعی و پیکره‌های تقویت شده در زبان فارسی مانند پیکره‌های وابستگی و یا معنایی زبان فارسی، استفاده از روش‌های مستقل از زبان نیز در این حوزه با چالش‌هایی همراه خواهد بود.

بنابراین بررسی روش‌های استخراج آزاد اطلاعات در زبان فارسی، و همچنین ارائه بستری برای استخراج با کیفیت اطلاعات در زبان فارسی و در دامنه وسیعی مانند وب، که این بستر شامل روش‌هایی برای پیدا کردن الگوهای ظهور اطلاعات در زبان فارسی، روش‌هایی برای یادگیری این الگوها و استخراج اطلاعات در زبان فارسی، روش‌هایی برای ابهام‌زدایی از موجودیت‌ها و روابط موجود در این اطلاعات، و نهایتا روش‌هایی برای ارزیابی میزان صحت اطلاعات استخراج شده، یکی از حوزه‌های بکر پژوهشی در زبان فارسی قلمداد می‌شو که می‌تواند مورد توجه پژوهشگران در حوزه پردازش زبان طبیعی و به ویژه استخراج اطلاعات قرار گیرد.

۶. مراجع

[1] A. Yates and O. Etzioni, “Unsupervised methods for determining object and relation synonyms on the web,” *J. Artif. Intell. Res.*, vol. 34, pp. 255–296, 2009.

[2] O. Etzioni, M. Banko, and M. Cafarella, “Machine Reading.,” *AAAI*, 2006.

[3] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, “YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia,” *Artif. Intell.*, vol. 194, pp. 28–61, Jan. 2013.

[4] S. Auer, C. Bizer, G. Kobilarov, and J. Lehmann, “Dbpedia: A nucleus for a web of open data,” *Semant. Web*, 2007.

[5] F. Peng and A. Mccallum, “Accurate Information Extraction from Research Papers using Conditional Random Fields,” 2003.

[6] D. Zelenko, C. Aone, and A. Richardella, “Kernel methods for relation extraction,” *J. Mach. Learn. …*, vol. 3, pp. 1083–1106, 2003.

[7] M. banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, *Open Information Extraction for the Web*. University of Washington, 2009.

[8] F. Wu and D. Weld, “Open information extraction using Wikipedia,” *Proc. 48th Annu. Meet. …*, no. July, pp. 118–127, 2010.

[9] A. Fader, S. Soderland, and O. Etzioni, “Identifying relations for open information extraction,” *Proc. Conf. …*, 2011.

[10] E. Agichtein and L. Gravano, “Snowball: Extracting relations from large plain-text collections,” *Proc. fifth ACM Conf. …*, no. December, 2000.

[11] O. Etzioni, M. Cafarella, and D. Downey, “Web-scale information extraction in knowitall:(preliminary results),” *Proc. 13th …*, 2004.

[12] M. Paşca, D. Lin, and J. Bigham, “Names and similarities on the web: fact extraction in the fast lane,” *Proc. 21st …*, no. July, pp. 809–816, 2006.

[13] M. Banko and O. Etzioni, “Strategies for lifelong knowledge extraction from the web,” *Proc. 4th Int. Conf. Knowl. capture - K-CAP ’07*, p. 95, 2007.

[14] A. Kulkarni and T. Pedersen, “Name Discrimination and Email Clustering using Unsupervised Clustering and Labeling of Similar Contexts.,” in *IICAI*, 2005, pp. 703–722.

[15] X. Li, P. Morie, and D. Roth, “Identification and tracing of ambiguous names: Discriminative and generative approaches,” in *Proceedings of the National Conference on Artificial Intelligence*, 2004, pp. 419–424.

[16] G. S. Mann and D. Yarowsky, “Unsupervised personal name disambiguation,” in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 2003, pp. 33–40.

[17] G. Salton and M. J. McGill, “Introduction to Moderm Information Retrieval.” McGraw-Hill, 1983.

[18] K. Rajaraman and a.-H. Tan, “Mining semantic networks for knowledge discovery,” *Third IEEE Int. Conf. Data Min.*, pp. 633–636, 2003.

[19] M. Lesk, “Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone,” in *Proceedings of the 5th annual international conference on Systems documentation*, 1986, pp. 24–26.

[20] G. A. Carpenter, S. Grossberg, and D. B. Rosen, “Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system,” *Neural networks*, vol. 4, no. 6, pp. 759–771, 1991.

[21] A. Haghighi and D. Klein, “Simple coreference resolution with rich syntactic and semantic features,” *Proc. 2009 Conf. Empir. …*, no. August, pp. 1152–1161, 2009.

[22] A. Haghighi and D. Klein, “Unsupervised coreference resolution in a nonparametric bayesian model,” in *Annual meeting-Association for Computational Linguistics*, 2007, vol. 45, no. 1, p. 848.

[23] T. Hasegawa, S. Sekine, and R. Grishman, “Discovering relations among named entities from large corpora,” *… 42nd Annu. Meet. …*, 2004.

[24] Y. Shinyama and S. Sekine, “Preemptive information extraction using unrestricted relation discovery,” *Proc. main Conf. Hum. …*, no. June, pp. 304–311, 2006.

[25] T. Lin and O. Etzioni, “Entity linking at web scale,” *… Knowl. Base Constr. Web-scale …*, 2012.

[26] R. Barzilay and L. Lee, “Learning to paraphrase: An unsupervised approach using multiple-sequence alignment,” in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 2003, pp. 16–23.

[27] Y. Shinyama and S. Sekine, “Paraphrase acquisition for information extraction,” in *Proceedings of the second international workshop on Paraphrasing-Volume 16*, 2003, pp. 65–71.

[28] B. Pang, K. Knight, and D. Marcu, “Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences,” in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 2003, pp. 102–109.

[29] W. B. Dolan and C. Brockett, “Automatically constructing a corpus of sentential paraphrases,” in *Proc. of IWP*, 2005.

[30] I. Dagan, O. Glickman, and B. Magnini, “The pascal recognising textual entailment challenge,” in in *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, Springer, 2006, pp. 177–190.

[31] S. Sekine, “Automatic paraphrase discovery based on context and keywords between ne pairs,” *Proc. IWP*, pp. 80–87, 2005.

[32] D. Davidov and A. Rappoport, “Unsupervised Discovery of Generic Relationships Using Pattern Clusters and its Evaluation by Automatically Generated SAT Analogy Questions.,” *ACL*, no. June, pp. 692–700, 2008.

[33] A. Yates and O. Etzioni, “Unsupervised Resolution of Objects and Relations on the Web.,” *HLT-NAACL*, no. April, pp. 121–130, 2007.

[34] D. Hindle, “Noun classification from predicate-argument structures,” in *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, 1990, pp. 268–275.

[35] A. E. Monge, C. Elkan, and others, “The Field Matching Problem: Algorithms and Applications.,” in *KDD*, 1996, pp. 267–270.

[36] W. W. Cohen, P. D. Ravikumar, S. E. Fienberg, and others, “A Comparison of String Distance Metrics for Name-Matching Tasks.,” in *IIWeb*, 2003, vol. 2003, pp. 73–78.

[37] D. Downey, O. Etzioni, and S. Soderland, “A probabilistic model of redundancy in information extraction,” 2006.

[38] S. Kok and P. Domingos, “Extracting Semantic Networks from Text Via Relational Clustering,” in in *Machine Learning and Knowledge Discovery in Databases*, 2008, pp. 624–639.

[39] M. Richardson and P. Domingos, “Markov logic networks,” *Mach. Learn.*, vol. 62, no. 1–2, pp. 107–136, 2006.

[40] S. Kok and P. Domingos, “Statistical predicate invention,” in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 433–440.

[41] A. McCallum and D. Jensen, “A note on the unification of information extraction and data mining using conditional-probability, relational models,” 2003.

[42] A. McCallum, K. Nigam, and L. H. Ungar, “Efficient clustering of high-dimensional data sets with application to reference matching,” in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 169–178.

[43] D. Lin and P. Pantel, “DIRT discovery of inference rules from text,” *… Conf. Knowl. Discov. data Min.*, pp. 0–5, 2001.

[44] A. Moro and R. Navigli, “WiSeNet: building a wikipedia-based semantic network with ontologized relations,” *Proc. 21st ACM Int. …*, pp. 1672–1676, 2012.

[45] A. Carlson, J. Betteridge, and B. Kisiel, “Toward an Architecture for Never-Ending Language Learning.,” *AAAI*, 2010.

[46] O. Etzioni, M. Cafarella, and D. Downey, “Unsupervised named-entity extraction from the web: An experimental study,” *Artif. Intell.*, pp. 1–42, 2005.

[47] G. Weikum and M. Theobald, “From information to knowledge: harvesting entities and relationships from web sources,” *Proc. twenty-ninth ACM SIGMOD …*, 2010.

[48] F. Suchanek, G. Kasneci, and G. Weikum, “Yago: a core of semantic knowledge,” *Proc. 16th …*, no. November, 2007.

[49] S. Ponzetto and M. Strube, “WikiTaxonomy: A Large Scale Knowledge Resource.,” *ECAI*, pp. 751–752, 2008.

[50] F. Suchanek, M. Sozio, and G. Weikum, “SOFIE: a self-organizing framework for information extraction,” *Proc. 18th …*, pp. 631–640, 2009.

[51] S. Schoenmackers, O. Etzioni, D. S. Weld, and J. Davis, “Learning first-order horn clauses from web text,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010, pp. 1088–1098.

[52] S. Schoenmackers, O. Etzioni, and D. Weld, “Scaling textual inference to the web,” *Proc. Conf. …*, no. October, pp. 79–88, 2008.

[53] W. C. Salmon, *Statistical explanation and statistical relevance*. University of Pittsburgh Pre, 1971.

[54] A. Fader, L. Zettlemoyer, and O. Etzioni, “Paraphrase-Driven Learning for Open Question Answering,” 2013.

[55] N. Balasubramanian, S. Soderland, O. Etzioni, and others, “Rel-grams: a probabilistic model of relations in text,” in *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, 2012, pp. 101–105.