مجله علمی-پژوهشی رایانش نرم و فن‌آوری اطلاعات

جلد ۵، شماره ۴، زمستان سال ۱۳۹۵ شاپا: ۱۰۰۶-۲۳۸۳

پیش‌بینی رویدادهای اخبار براساس استنتاج علّی در منطق مرتبه اول

سینا دامی*[۱]، حسین شیرازی[۲]، احمد عبداله‌زاده بارفروش[۳]

*۱- نویسنده مسئول: استادیار، مجتمع دانشگاهی فناوری اطلاعات ارتباطات و امنیت، دانشگاه صنعتی مالک اشتر، تهران، ایران، Dami@mut.ac.ir

۲- دانشیار، مجتمع دانشگاهی فناوری اطلاعات ارتباطات و امنیت، دانشگاه صنعتی مالک اشتر، تهران، ایران، Shirazi@mut.ac.ir

۲- استاد، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیر کبیر، تهران، ایران، Ahmad@ce.aut.ac.ir

چکیده- در این مقاله یک روش جدید برای پیش‌بینی رویدادهای اخبار در محیط‌های متنی ارائه شده است. روش پیشنهادی، از طریق تعمیم رویدادهای علت و سپس پیش‌بینی رویدادهای معلول آن با استفاده از قواعد علّی، قادر به تولید مدل پیش‌بینی رویداد است. برای این منظور، ابتدا رویدادهای موردنظر ازطریق یک مدل بازنمایی رویداد در سطح معنا از پیکره متنی اخبار استخراج شده و در قالب یک مدل گرافیکی در آنتولوژی (پایگاه شناخت) به عنوان دانش پویا ذخیره می‌شود. سپس یک سری قواعد علّی خاص حوزه پیش‌بینی در قالب منطق مرتبه اول به ماشین تزریق می‌گردد. برای مدل کردن دانش ایستا علاوه بر پایگاه قواعد، از چندین پایگاه شناخت بزرگ شامل پایگاه شناخت عمومی نظیر DBpedia، پایگاه شناخت واژگان نظیر FarsNet و پایگاه شناخت افعال نظیر VerbNet، به عنوان دانش ضمنی برای تعمیم دادن رویدادها و تولید مدل پیش‌بینی استفاده می‌شود. در نهایت، تمامی این مدل‌ها در قالب استاندارد زبان پایگاه شناخت وب (OWL)، به منظور انجام استنتاج علّی تجمیع می‌شوند. ارزیابی تجربی و عملی در اخبار واقعی نشان داد، که روش پیشنهادی برای پیش‌بینی اخبار عملکرد بهتری نسبت به روش‌های پایه دارد.

واژه‌های کلیدی: پیش‌بینی اخبار، بازنمایی رویداد، پردازش معنایی، آنتولوژی، استنتاج علّی، منطق مرتبه اول.

۱- مقدمه

درگذشته عموماً پیش‌بینی رویدادها بر اساس اخبار توسط خبرگان این زمینه انجام می‌شد. برای مثال خبرگانی در زمینه اخبار سیاسی با توجه به انعکاس رخدادهای قبلی در اخبار حوادث و رخدادهای آتی را پیش‌بینی می‌کردند. در واقع روابط علّی بین رخدادها در این پیش‌بینی‌ها مورد استفاده قرار می‌گیرد. خبرگان این زمینه با یادگیری این روابط در طول حرفه خود این روابط را در پیش‌بینی رخدادها استفاده می‌کنند.

پژوهش‌های روانشناسی نشان می‌دهد [۱] انسان‌ها قابلیت بالایی در درک مفاهیم علّی بین رخدادها دارند. از بدو تولد روابط بین رخدادها در مغز انسان شکل می‌گیرد. در آینده انسان از این روابط بهره گرفته و در پیش‌بینی رخدادها به کار می‌بندد. انسان‌ها با استفاده از تجربیات می‌توانند رفتار موجودیت‌های مختلف در طبیعت را مدل‌سازی و پیش‌بینی کنند. این مطالعات شواهدی را نشان می‌دهد که انسان‌ها در انواع خاصی از پیش‌بینی رویدادها بسیار خوب عمل می‌کنند. به‌خصوص هنگامی که آن‌ها براساس دانشی که در طول عمرشان درباره نحوه رفتار جهان به‌دست آوردند، عمل نمایند. برای این منظور، آن‌ها به طور گسترده یک سری دانش عام[1] در اختیار دارند: جهان در حال حاضر چگونه است؟ قبل از آن چگونه بود؟ چه اتفاقی بعد از یک عمل می‌افتد؟ چه رویدادی موجب دیگری می‌شود؟

حال سوال اساسی این است که آیا می‌توان چارچوبی برای ماشین‌ها طراحی کرد که بتوانند قابلیتی همچون انسان در پیش‌بنی رویدادها داشته باشند؟ به عبارت دیگر آیا می‌توان قواعد علّی بین رویدادها را در ماشین پیاده‌سازی کرد؟ چگونه می‌توان با استفاده از دانش عام و ترکیب انواع مختلف دانش عمل پیش‌بینی رویداد را انجام داد؟ چگونه می‌توان داده‌هایی را که از منابع اطلاعاتی متعدد و ناهمگنی می‌آید، با هم تجمیع نمود و بر روی داده‌های ناکامل عمل استنتاج انجام داد؟ چگونه می‌توان یک رویداد را به صورت یک مدل مستقل از زبان[2] بازنمایی نمود، به

[1] Common-sense knowledge
[2] Language-independent

طوری‌که بتوان عمل پیش‌بینی را بر روی آن انجام داد؟ قابلیتی که ماشین ورای انسان دارد امکان بررسی و تعامل با حجم بسیار زیادی از داده‌ها است. شبکه جهانی اینترنت این امکان را فراهم می‌سازد که ماشین حجم بسیار زیادی از اخبار گذشته را همراه با تاثیر هرکدام در رویدادها بررسی کرده و مدلی برای روابط علّی رویدادها بسازد. هم‌چنین پایگاه‌های شناخت[1] بسیاری در شبکه جهانی موجود است که با استفاده از آن می‌توان روابط علّی رویدادها را بررسی کرد. علاوه بر این استفاده از مزایای منطق برای بازنمایی رویدادها می‌تواند راه‌حلی برای پیچیدگی مسئله ارائه دهد.

با وجود مطالعات زیادی که در حوزه استخراج اطلاعات از متون انجام شده است [۲, ۳]، مطالعات کمی در حوزه استخراج روابط علّی صورت گرفته شده است ولی کارهای [۴, ۵] استثناهای قابل توجه‌ای هستند. گذشته از این، الگوریتم‌های توسعه یافته شده برای استخراج روابط علّی سعی در شناسایی این روابط دارند و به این دلیل که قادر به تولید رویدادهای جدید ناشی از یک رویداد ممکن نیستند، نمی‌توانند برای پیش‌بینی مورد استفاده قرار بگیرند.

از آنجایی که مسأله پیش‌بینی آن هم در زمینه اخبار مبهم و گاه پیچیده است، تعریف دقیق این مسأله برای مشخص شدن رویکرد و راه‌حل بسیار اهمیت دارد. هرگاه صحبت از اخبار می‌شود ذهن انسان متوجه نوشته‌های روزنامه در مورد رویدادهای اخیر خواهد بود. در این مسأله به دنبال پیش‌بینی دقیق این نوشته‌ها که در واقع بازخوردی از رویدادها هستند، نیستیم. بلکه در عوض به دنبال پیش‌بینی رویدادهایی هستیم که در آینده واقع خواهند شد و در اخبار نیز مشاهده می‌شوند.

به طور دقیق‌تر داده‌هایی که در اختیار داریم پیکره‌ای از اخبار روزهای گذشته است. ماشین باید با در نظر گرفتن این اخبار و در واقع رخدادهایی که از این اخبار استنتاج می‌شود، رخدادهای آتی را پیش‌بینی کند. یکی از قسمت‌های اساسی در طراحی و پیاده‌سازی این ماشین بخش پردازش زبان طبیعی است. رویکردهای متفاوتی برای استخراج مفاهیم از زبان طبیعی ارائه شده است [۶]. برخی رویکردها با توجه به تعداد تکرار کلمات مهم در هر حوزه ویژگی‌هایی به عنوان مفهوم از متن استخراج می‌کنند. این رویکرد انباشت واژگان[2] نامیده می‌شود و در بسیاری از حوزه‌ها مانند طبقه‌بندی متون و تصاویر عملکرد مناسبی داشته است. در این مدل‌ها توجه کمی به مفاهیم می‌شود و در واقع ویژگی‌های استخراج شده فاصله زیادی با مفاهیم موجود در متن دارد. برای مثال در این روش‌ها جمله‌های «عراق به ایران حمله کرد» و «ایران به عراق حمله کرد» ویژگی‌های یکسانی دارند و به عنوان یک داده تلقی می‌شوند. رویکرد دیگر استفاده از نقش‌های ساختاری کلمات است. روش‌های متعددی برای تشخیص ماهیت ساختاری کلمات موجود است. این روش‌ها می‌توانند یک عبارت اسمی را از فعل و یا صفت متمایز کنند. به این ترتیب می‌توان در استخراج ویژگی‌ها ساختار کلمات را نیز در نظر گرفت. حتی با در نظر گرفتن ویژگی‌های استخراج شده با این روش دو جمله مثال قبلی از یکدیگر قابل تمایز نیستند. رویکردی که می‌تواند برای تشخیص مفاهیم در جملات خبری مورد استفاده قرار گیرد، استفاده از نقش‌های مفهمومی کلمات در جمله است [۷]. مثال‌هایی از این نقش‌ها در جملات فارسی فاعل، مفعول و نقش‌هایی از این قبیل است. طبقه‌بندی‌های زیادی برای این نقش‌ها در جملات انگلیسی موجود است. این نقش‌ها معمولاً با نام نقش‌های موضوعی[3] شناخته می‌شوند. استفاده از این نقش‌ها مستلزم ارائه روشی خودکار برای استخراج این نقش‌ها از جملات است. روش‌های موفقی برای این مسأله با استفاده از یادگیری ماشین ارائه شده است [۸, ۹].

روش‌های متعدد دیگری نیز برای پیش‌بینی رویداد با توجه به بازخوردهای موجود در شبکه‌های اجتماعی ارائه شده است [۱۰, ۱۱]. معمولا همزمان با تحولات اجتماعی کاربران شبکه‌های اجتماعی مثل توئیتر و فیس‌بوک این تحولات را در نوشته‌های خود منعکس می‌کنند. این نوشته‌ها هم‌چنین می‌تواند برای پیش‌بینی وقایع آتی مورد استفاده قرار گیرد. نکته قابل توجه در این رویکرد حجم بالایی از متون کوتاه در مورد مسائل اجتماعی است. در [۱۲] بررسی و مطالعه‌ای به طور خاص در مورد پیش‌بینی فروش فیلم‌ها در گیشه یا همان پیش‌بینی درآمد حاصل از فروش بلیط برای فیلم‌ها با استفاده از چاترهای توئیتر انجام گرفته است. در این پژوهش نشان داده شد که چگونه می‌توان از رسانه‌های اجتماعی برای پیش‌بینی آینده استفاده نمود. با استفاده از سه میلیون توئیت از سایت توئیتر، یک مدل رگرسیون خطی برای پیش‌بینی درآمد فیلم‌ها پس از انتشار ساخته شد. همچنین نشان داده شد که این روش دقت مناسبی برای پیش‌بینی بورس و اوراق بهادار هالیوود دارد و یک هم‌سانی بسیار قوی بین رتبه فیلم در آینده و تعداد توجهات آن وجود دارد. همچنین با استفاده از تحلیل احساسات[4] نشان داده شد که این نوع از تحلیل باعث بهبود

[1] Ontologies
[2] Bag of words
[3] Thematic Roles
[4] Sentiment analysis

پیش‌بینی خواهد شد. اگر چه در این مطالعه بر روی پیش‌بینی درآمد فیلم تمرکز شده است، اما می‌توان گفت که این روش می‌تواند برای سایر محصولات و علایق مصرف‌کنندگان در محدوده وسیعی از پیش‌بینی‌ها، از رتبه‌بندی محصولات مختلف در آینده گرفته تا تنظیم دستورکار و نتایج انتخابات به کارگرفته شود. در روابط عمیق‌تر این کار نشان می‌دهد رسانه‌های اجتماعی بیانگر خرد جمعی است که اگر به درستی مورد بررسی قرار بگیرند، می‌توانند منجر به پیش‌بینی بسیار دقیق نتایج آینده شوند.

مسائل مشابهی نیز مورد توجه پژوهش‌گران قرار گرفته است که بسیار مشابه پیش‌بینی رویدادها هستند. پیش‌بینی قیمت سهام و یا فلزات گران‌بها یکی از مسائل مشابه است [۱۳, ۱۴]. در واقع رویداد خاصی در این مسأله مورد نظر است و هدف از طراحی الگوریتم پیش‌بینی وقوع این رویداد خاص است. از آنجا که رسانه‌های اجتماعی می‌توانند به عنوان یک شکل از خرد جمعی تعبیر شوند، قدرت آن‌ها در پیش‌بینی نتایج در دنیای واقعی بررسی شده است. با کمال تعجب ملاحظه می‌شود که در واقع یک جامعه مجازی می‌تواند در پیش‌بینی‌های کمّی بهتری، نسبت به بازار مجازی مورد استفاده قرار گیرد. این بازارهای اطلاعاتی به طور کلی معاملات دولتی مشروط اوراق بهادار را شامل می‌شود و اگر به اندازه کافی بزرگ و به درستی طراحی شده باشد، معمولاً دقیق‌تر از روش‌های دیگر برای استخراج اطلاعات منتشر شده مانند نظرسنجی‌ها هستند. به طور خاص نشان داده شده است که قیمت‌ها در این بازارها، دارای ارتباط قوی با فرکانس نتایج مشاهده شده است [۱۵, ۱۶] و در نتیجه شاخص‌های خوبی برای نتایج آینده هستند. علاوه بر این، جمع‌آوری اطلاعات در مورد نحوه رفتار مردم نسبت به یک محصول خاص، در هنگام طراحی کمپین‌های بازاریابی و تبلیغاتی یکی از وظیفه‌های مهم است [۱۷, ۱۸].

یکی دیگر از مسائل مطرح، پیش‌بینی درخواست کاربران از طریق جست‌وجو در شبکه اینترنت است [۱۹]. تعدادی از پژوهش‌های اخیر بر روی بررسی رابطه بین پرس‌وجوهای موجود در وب و رویدادهای اخبار تمرکز داشته‌اند. رادینسکی و همکاران [۲۰] نشان دادند که پرس‌وجوهای کاربران در وب می‌تواند بازتابی در رویدادهای آینده اخبار داشته باشد. گینسبِرگ و همکاران [۲۱] نیز از پرس‌وجوها برای پیش‌بینی شیوع آنفولانزای H1N1 استفاده کرده‌اند. روش دیگری که در [۱۰] استفاده شده است، استفاده از روش‌های پردازش سیگنال و فیزیک به منظور مدل‌سازی و پیش‌بینی تغییرات رفتاری در وب است. در این شیوه از مدل حالت-فضا[1] [۲۲] به عنوان یکی از روش‌های معمول برای بررسی سیستم‌های پویا در فیزیک و پردازش سیگنال استفاده شده است. در مطالعه دیگری [۱۱] از مدل‌های سری زمانی[2] برای پیش‌بینی رویدادهای آتی بر اساس داده‌های متنی ارایه شده است. این مدل بر اساس سیستم‌های بازیابی اطلاعات[3] موجود عمل می‌کند. به این ترتیب که کاربر کلمه‌ای را به عنوان پرس‌وجو درنظر می‌گیرد و سیستم بازیابی تمام متون مرتبط با کلمه را جدا می‌کند. سری‌های زمانی بر اساس کاربرد کلمه مورد پرس‌وجو در متون شکل می‌گیرد و مدل استفاده شده در کنار یک مدل احتمالاتی[4]، احتمال وقوع رویدادها را پیش‌بینی می‌کند. از مزیت‌های این روش، استفاده از پرس‌وجوی کاربران به‌صورت سری‌های زمانی است که در مدل‌های پیشین به آن توجه نشده است.

فعالیتی که ما قصد انجام آن را داریم از جهات مختلف با مطالعات ذکر شده متفاوت است. ابتدا اینکه، ما قصد ارائه یک مدل معنایی و مستقل از زبان را برای پیش‌بینی رویدادها داریم. دوم اینکه، برخلاف مطالعات ذکر شده، ما از منابع اطلاعاتی ناهمگنی (نظیر Wordnet، Wikipedia و ...) برای پیش‌بینی استفاده خواهیم کرد. در آخر اینکه، مطالعه ما برروی تولید پیش‌بینی‌هایی از رویدادهای آینده تمرکز خواهد داشت که کاملاً در قالب زبان طبیعی باشند، و همچنین روش‌هایی برای غنی‌سازی و تعمیم دادن رویدادهای استخراج شده به منظور پیش‌بینی رویدادهای آینده ارائه خواهیم کرد.

هدف ما در این پژوهش ارائه رویکردهایی برای انجام استنتاج علّی[5] به منظور پیش‌بینی رویدادهای معلول در محیط‌های متنی است. در حالت کلی، کار ما بر روی بهبود روش‌های استخراج اطلاعات[6] یا استخراج روابط علّی[7] خلاصه نمی‌شود، بلکه بر روی چگونگی و نحوه‌ی به‌کارگیری این اطلاعات برای امر پیش‌بینی تمرکز دارد. ما سعی در ارائه روشی جدید برای ترکیب دانش هستی با روش‌های استخراج رویداد جهت بازنمایی رویدادهای منسجم و همچنین ارائه‌ی یک مدل جدید با استفاده از روش‌های استنتاج علّی در منطق مرتبه اول[8] (FOL) جهت تعمیم[9] رویدادهای استخراج شده با استفاده از این دانش هستیم. به دلیل اهمیت و گستردگی بحث پایگاه شناخت (آنتولوژی) و ساخت و

[1] State-Space Model
[2] Time-Series Model
[3] Information Retrieval
[4] Probabilistic Model
[5] Causality reasoning
[6] Information extraction
[7] Causality extraction
[8] First-Order Logic
[9] Generalization

استفاده از آن در پیش‌بینی رویدادهای آتی، در بخش ۲ به طور مجزا به این موضوع پرداخته می‌شود. در بخش ۳ طرح مسأله و راه‌حل پیشنهادی ارائه می‌شود. در بخش ۴ نحوه ارزیابی پیش‌بینی به همراه نتایج آن بیان می‌شود. در نهایت، نتیجه‌گیری و پیشنهاد برای پژوهش‌های آتی در بخش ۵ تشریح می‌گردد.

۲- پایگاه شناخت

یکی از قسمت‌های عمده این پژوهش پیاده‌سازی و کار با انواع پایگاه شناخت‌ها است. همچنین کلیه دانش به دست آمده در پایگاه شناخت ذخیره شده و استنتاج نیز با توجه به آن و در قالب آن انجام می‌شود. به همین دلیل استخراج پایگاه شناخت، بازنمایی و استنتاج آن اهمیت فراوانی دارد.

۲-۱- استخراج پایگاه شناخت

استخراج پایگاه شناخت نوع خاصی از استخراج اطلاعات است. هدف از استخراج پایگاه شناخت، ساخت سلسله مراتبی از مفاهیم و معمول‌ترین روابط بین آن‌هاست. مطالعات در زمینه استخراج پایگاه شناخت به طور عمده بر روی استخراج دسته‌بندی از انواع مختلفی از منابع تاکید دارد. یک دسته‌بندی، یک گراف جهت‌دار بدون‌دور است که در آن مفاهیم با روابطی نظیر IS-A به یکدیگر متصل شده‌اند.

با توجه به آزمایشات انجام شده در پژوهش‌های پیشین [۲۳-۲۵]، استفاده از پایگاه شناخت‌های بزرگ نیازمند پردازش‌هایی است که توسط یک کامپیوتر معمولی امکان‌پذیر نمی‌باشد. و ممکن است، نیاز به پردازش‌های موازی یا بستر توزیع‌شده وجود داشته باشد. در این پژوهش، برای رفع این مشکل یک واسط برای کار با پایگاه شناخت‌ها در نظر گرفته شده است، که وظیفه‌ی بارگذاری، خلاصه‌سازی و پردازش در پایگاه شناخت‌های ورودی و خروجی را بر عهده دارد. از این‌رو ایجاد یک بستر موازی و توزیع‌شده برای کار با پایگاه شناخت و انجام استنتاج به عنوان یکی از کارهای آتی رها می‌شود.

۲-۲- بازنمایی پایگاه شناخت

پیش‌نیاز کاربرد پایگاه شناخت‌ها در وب معنایی، توسعه استانداردی برای تعریف و مبادله پایگاه شناخت یا به عبارتی زبان‌های بازنمایی پایگاه شناخت است. مهمترین استانداردهای کنونی برای بازنمایی پایگاه شناخت‌ها، اطلاعات را به‌صورت داده‌های پیوندی[1] بیان می‌کنند. در این راستا کنسرسیوم وب جهانی (W3C)، زبان چارچوب توصیف منابع (RDF) را توسعه داد. RDF، زبانی برای کدگذاری دانش موجود در صفحات وب، به منظور قابل فهم کردن این دانش برای عامل‌های الکترونیکی جستجوگر اطلاعات بود. موسسه DARPA نیز با همکاری W3C، زبان DAML را ایجاد کرد. زبان DAML گسترشی از RDF بود که در آن ساخت‌های توصیفی بیشتری استفاده می‌شد. هدف این ساخت‌ها تسهیل تعامل عامل‌ها در وب بود. سرانجام زبان OWL بر مبنای زبان DAML ایجاد شد. استاندارد OWL یک زبان بازنمایی بر مبنای منطق توصیفی[2] است، که امکان بیان روابط حاکم بین مجموعه‌ها را داراست. از میان زبان‌های مطرح شده، OWL جدیدترین استاندارد زبانی توصیه شده است و قابلیت‌های بیشتری نیز نسبت به سایر زبان‌ها دارد. به‌کارگیری زبان‌های بازنمایی دانش، امکان تعریف، برقراری ارتباط میان پایگاه شناخت‌ها و تحقق کنش‌پذیری و سازگاری با هم را فراهم می‌کند.

۲-۳- استنتاج در پایگاه شناخت

پیش فرض‌هایی برای استنتاج در زبان OWL قرار داده شده است. برای مثال می‌توان پرسید که آیا نمونه‌ای در دسته خاص حضور دارد یا خیر. و یا اینکه دسته‌ها زیر دسته یکدیگر هستند یا نه. با این حال برای برای بیان قواعد معنایی و استنتاج بر اساس آن نیاز به ماشین استنتاج قواعد با پشتیبانی از قواعد خاص پیچیده‌تر مثل منطق مرتبه اول داریم. ابزارهایی در کنار استاندارد OWL معرفی شده‌اند که این قابلیت را در کنار ماشین استنتاج خود به این زبان افزوده‌اند. برای مثال کتابخانه Jena[3] توانایی گرفتن قواعد را در قالب منطق مرتبه اول دارد. علاوه بر این، در این کتابخانه امکان بارگذاری پایگاه شناخت‌های ذخیره شده در قالب‌های مختلف نیز وجود دارد. بدین منظور برای پیاده‌سازی روش ارائه شده از این ابزار استفاده شده است.

یک پایگاه دانش مرتبه اول[4]، مجموعه‌ای از عبارات یا فرمول‌ها در منطق مرتبه اول است [۲۶]. فرمول‌ها با استفاده از چهار نوع علامت ساخته شده‌اند: ثابت‌ها، متغیرها، توابع و مسندها[5]. یک تفسیر، مشخص می‌کند که کدام موضوعات، توابع و رابطه‌ها در حوزه به وسیله‌ی کدام علامت نمایش داده می‌شوند. متغیرها و ثابت‌ها ممکن است طبقه‌بندی باشند؛ متغیرهای نمونه، تنها روی موضوعات مدل متناظر تنظیم می‌شوند و ثابت‌ها فقط می‌توانند موضوعات مدل متناظر را نمایش دهند. برای مثال، متغیر x ممکن

[1] LinkedData

[2] Description Logic
[3] http://jena.apache.org/
[4] First-order knowledge base
[5] Predicates

است روی افراد (مانند Anna، Bob و ...) تنظیم شود و ثابت C یک شهر (مانند Seattle، Tokyo و.....) را نمایش می‌دهد. هر جمله در هر عبارت، یک موضوع در حوزه را نشان می‌دهد. یک جمله می‌تواند، یک ثابت، یک متغیر یا یک تابع باشد که به یک چندتایی جملات اعمال شده باشد: به‌عنوان مثال، Anna، X و GreatestCommonDivisor(x, y) جمله هستند. یک فرمول اتمی یا اتم، یک علامت مسند است که به یک چندتایی جملات (مانند Friend(x, MotherOf(Anna))) اعمال شده است. فرمول‌ها به صورت بازگشتی با استفاده از سورها و رابطهای منطقی از فرمول‌های اتمی ساخته شده‌اند. یک عبارت گراوند[1]، عبارتی است که هیچ متغیری را دربر نمی‌گیرد، یک اتم گراوند یا مسند گراوند یک فرمول اتمی است که تمام شناسه‌های آن جملات گراوند می‌باشد. یک مجموعه ممکن (به همراه یک تفسیر) یک مقدار صحیح را به هر اتم پایه ممکن اختصاص می‌دهد.

یک فرمول F، ارضا شدنی[2] است اگر و تنها اگر حداقل یک مجموعه صحیح در آن وجود داشته باشد. مسئله اصلی استنتاج در منطق مرتبه اول تعیین کردن این است که آیا یک پایگاه دانش KB، یک فرمول F را استلزام[3] می‌کند. اگر F در همه جا صحیح باشد، آنجا KB نیز صحیح است (به صورت KB |= F نوشته می‌شود). این اغلب از طریق رد قضیه[4] انجام می‌شود: F مستلزم KB است، اگر و تنها اگر KB∪¬F غیرقابل ارضا شدن باشد. (بنابراین اگر یک KB شامل یک تناقض باشد، تمام فرمول‌ها به‌طور جزئی از آن متابعت می‌کنند). برای انجام عمل استنتاج خودکار، اغلب به فرمول‌هایی تبدیل می‌شوند که دارای شکل منظم‌تری باشند، به‌طور نمونه به فرم بند[5] (که به فرم نرمال عطفی (CNF) نیز شناخته شده‌اند). یک KB در فرم بند، یک ترکیب عطفی از بندها است، یک بند نیز ترکیب فصلی از لفظها می‌باشد. هر KB در منطق مرتبه اول می‌تواند با استفاده از یک توالی ماشینی از مراحل به فرم بند تبدیل شود (این تبدیل، شامل حذف سور وجودی توسط Skolemnization است که در حالت کلی درست نمی‌باشد). فرم بند در یک روال استنتاج درست[6]، تفکیک‌پذیر[7] و رد کامل[8] برای منطق مرتبه اول مورد استفاده قرار می‌گیرد.

[1] Ground term
[2] Satisfiable
[3] Entails
[4] Refutation
[5] Clausal form
[6] Sound
[7] Resolution
[8] Refutaton-complete

۲-۴- پایگاه شناخت‌های بررسی شده

امروزه تعداد بسیار زیادی پایگاه شناخت برای پروژه‌های تحقیقاتی طراحی شده و مورد استفاده قرار می‌گیرد. در [۲۳] از یکی از بزرگترین پایگاه شناخت‌های موجود[9] که شامل بیش از یک میلیارد قاعده است استفاده شده است. برای استفاده از چنین پایگاه شناخت‌های با حجم بالا نیاز به سرورهای قدرتمند و پردازش موازی است. اگر بخواهیم به شکل عملی در مورد پایگاه شناخت‌ها بحث کنیم باید آن‌ها را از لحاظ بزرگی (تعداد قواعد موجود)، حجم داده ذخیره شده آن‌ها و هم‌چنین حافظه فیزیکی مورد نیاز بررسی کنیم. در جدول (۱) مشخصات پایگاه شناخت‌های بررسی شده آورده شده است. از آنجایی که در این پژوهش نیاز به دانش عمومی داریم، پایگاه شناخت‌هایی که برای زمینه‌های خاص طراحی شده‌اند، نمی‌توانند مورد استفاده قرار بگیرند و لذا در این مقایسه ذکر نشده است.

با وجود پایگاه شناخت‌های عمومی زیاد، تعداد کمی از این پایگاه شناخت‌ها موجودیت‌های زبان فارسی را مورد پوشش قرار می‌دهند. با توجه به این که این منابع در زبان انگلیسی بسیار غنی‌تر از زبان فارسی هستند، باید به دنبال راه‌حلی برای انتقال دانش موجود در آن‌ها به زبان فارسی باشیم. در بخش ۲-۵ به ارائه راه‌کاری در این خصوص پرداخته می‌شود. یکی از معروف‌ترین و در عین حال بزرگترین پروژه‌هایی که می‌توان به انتقال دانش آن به زبان فارسی امید داشت، پایگاه شناخت DBpedia [۲۷] است، زیرا شامل روابط سه‌تایی است، که مقدار آن‌ها اسامی ویژگی‌ها و مقدارهاست، و از طریق ترجمه آسیب نمی‌بیند. از این‌رو این پایگاه شناخت به علت جامع بودن، پوشش مناسب زبان فارسی، پشتیبانی از استانداردهای اخیر پایگاه شناخت نظیر OWL و حجم داده مناسب برای این پژوهش برگزیده شد.

در این پروژه اطلاعات متنوعی که در صفحات ویکی‌پدیا موجود است استخراج و طبقه‌بندی شده است. یکی از قسمت‌های پروژه DBPedia پایگاه شناخت عمومی است، که توسط جعبه‌های اطلاعاتی که در کنار صفحات ویکی قرار می‌گیرد، تولید شده است. این جعبه‌ها شامل یک سری ویژگی‌ها به همراه مقادیر مربوط به آن ویژگی‌ها هستند. این ویژگی‌ها در قالب OWL در پایگاه شناخت ذخیره شده‌اند. قسمت دیگری از پایگاه شناخت درباره دسته‌هایی است که موجودیت‌ها به آن تعلق دارند.

در چارچوب پیشنهادی – که در بخش ۳-۱ تشریح شده است - کاربر می‌تواند پایگاه شناخت دلخواه خود را با توجه به حوزه مورد

[9] Billion Triples Challenge Dataset 2009

بحث به سیستم اضافه کند. پایگاه شناخت می‌تواند در قالب فایل‌های استاندارد ذخیره شده باشد و یا از آدرسی در شبکه به سیستم اضافه شود. فایل‌های پشتیبانی شده در این سیستم عبارتند از: Turtle، RDF/XML، N-Triples، RDF/JSON، TriG و N-Quads.

علاوه بر پایگاه شناخت عمومی، کاربر می‌تواند ساختار پایگاه شناخت افعال و هم‌چنین پایگاه شناخت ورودی اخبار را نیز تغییر دهد. برای مثال اگر در بخش استخراج حقایق، نقش‌های موضوعی استخراج‌شده کم یا زیاد شود، این تغییرات می‌تواند در پایگاه شناخت اصلی سیستم پیش‌بینی درج شده و پیش‌بینی رویدادها با توجه به نقش‌های موضوعی جدید انجام شود.

حجم داده‌های موجود در پایگاه شناخت DBpedia بسیار بالا است و از طرف دیگر تمام این داده‌ها مورد استفاده سیستم مورد نظر قرار نمی‌گیرد. از این‌رو ماژول‌های برای تعمیم و تلخیص این پایگاه شناخت در نظر گرفته شده است، که به‌ترتیب در بخش‌های ۲-۶ و ۲-۷ تشریح می‌شود.

جدول ۱: بررسی پایگاه شناخت‌های عمومی.

نام پایگاه شناخت	حجم پایگاه شناخت (تعداد قواعد موجود)	حافظه موردنیاز (GB)
Billion Triples Challenge Dataset 2009	1.14 billion	500+
Billion Triples Challenge Dataset 2010	3.2 billion	1000+
Data-gov Wiki	5+ billion	1000+
DBpedia	247 million	500
Freebase RDF Store	505 Mbytes	500+
OpenCyc	~1.6 million	10
YAGO	1Gb	1000 +

۲-۵- یافتن نمونه‌های فارسی مناسب برای نمونه‌های انگلیسی پایگاه شناخت

یکی از چالش‌های اساسی در این پژوهش استفاده از پایگاه شناخت حجیم پروژه DBpedia و استفاده از آن در چارچوب فارسی است. تمامی نمونه‌ها[1]، دسته‌ها[2] و ویژگی‌ها[3] حاوی آدرس‌های وب در زبان انگلیسی است. دو روش کلی در بخش پیش‌بینی برای تبدیل موجودیت‌های این پایگاه شناخت به فارسی در نظر گرفته شده است.

رویکرد اول استفاده از یکی از قسمت‌های پروژه DBpedia است. در این پروژه صفحه‌های Wikipedia در زبان‌های مختلف در یک فایل حجیم پایگاه شناخت معادل‌سازی شده‌اند. برای مثال صفحه مربوط به تهران مربوط به دو لینک فارسی و انگلیسی در این فایل با یکدیگر به شکل زیر یکسان شده‌اند.

http://en.wikipedia.org/wiki/Tehran
http://fa.wikipedia.org/wiki/تهران

رویکرد دوم استفاده از مترجم گوگل است. موجودیت‌هایی که در قسمت قبل در صفحات فارسی Wikipedia یافت نشده است در این بخش توسط مترجم گوگل ترجمه می‌شود. یکی از مزیت‌های این رویکرد ترجمه اسم‌های خاص در مترجم گوگل است. برای مثال موجودیت "Sari" در مترجم گوگل تبدیل به "ساری" می‌شود، در حالی‌که مترجم‌های دیگر عموماً شامل اسم‌های خاص در زبان‌های مختلف نمی‌باشند. در این پژوهش از هر دو رویکرد فوق برای این منظور استفاده شده است.

۲-۶- تعمیم نمونه‌ها در پایگاه شناخت

همان‌طور که در بخش ۲-۴ اشاره شد، از DBpedia به عنوان پایگاه شناخت عمومی در سیستم پیشنهادی بهره گرفته شده است. پروژه DBpedia یک پایگاه شناخت چند زبانه با مقیاس بزرگ را با استخراج داده‌های ساخت‌یافته از نسخه‌های ویکی‌پدیا در ۱۱۱ زبان ساخته است. این پایگاه شناخت می‌تواند برای پاسخ به درخواست‌های توصیفی برخلاف آنچه که ویکی‌پدیا از آن محروم است، بکار رود. این پایگاه شناخت، ساختار چارچوب‌های متفاوت را به صورت مجزا هم در نسخه‌های زبانی ویکی‌پدیا و هم از میان ۲۷ زبان مختلف موجود یکی می‌کند.

پایگاه شناخت DBpedia به‌طور گسترده به‌عنوان یک بستر آزمایشی در جامعه پژوهشی بکار رفته است و کاربردها، الگوریتم‌ها و ابزارهای متعددی در حوزه DBpedia تولید شده یا به آن اعمال شده‌اند. به دلیل شمای بزرگ و اندازه داده و تنوع موضوعی آن، DBpedia چالش‌های علمی زیادی را ایجاد می‌کند. به خصوص تولید سیستم‌های توانمند در این حوزه برای DBpedia صرفا مبتنی بر الگوهای ساده یا از طریق لغت‌نامه‌های با دامنه خاص به دلیل اندازه و پوشش گسترده‌اش مشکل است. بنابراین، یک سیستم پیش‌بینی که بتواند به درستی جواب مطمئن از طریق استنتاج در DBpedia بدهد، می‌تواند به عنوان یک سیستم هوشمند واقعی لحاظ گردد.

برای این منظور و به جهت افزایش کارایی استفاده در DBpedia،

[1] Instances
[2] Classes
[3] Properties

فایل مربوط به نمونه‌ها و ویژگی‌های موجود در آن را مورد پردازش قرار دادیم. هدف از پردازش‌های انجام شده دراین بخش، در ادامه با مثالی توضیح داده شده است.

فرض کنید یک سه‌تایی به صورت زیر در فایل نمونه‌ها در DBpedia موجود باشد:

<http://dbpedia.org/resource/Tehran>
<http://dbpedia.org/ontology/capital>
<http://dbpedia.org/resource/ Iran>

در این صورت برنامه بایستی با توجه به هر نمونه مشابه نمونه فوق، فایل ویژگی‌ها در DBpedia را برای درایه اول جستجو کند و هر نوعی که بیشترین تکرار را برای این درایه داشته باشد، به عنوان *نوع* برای این درایه در نظر بگیرد. به عنوان مثال فرض کنیم بیشترین تکرار نوع برای درایه اول به صورت زیر باشد:

<http://dbpedia.org/resource/Tehran>
<http://www.w3.org/rdf-syntax-ns#type>
<http://dbpedia.org/ontology/City>

برای درایه سوم هم فایل ویژگی‌ها را مورد جستجو قرار می دهیم و بیشترین تکرار نوع برای این درایه را می‌یابیم. فرض کنیم بیشترین تکرار را عضو زیر داشته باشد:

<http://dbpedia.org/resource/Iran>
<http://www.w3.org/rdf-syntax-ns#type>
<http://dbpedia.org/ontology/Country>

در نهایت یک قانون به صورت زیر ایجاد می‌شود:

<http://dbpedia.org/resource/City>
<http://dbpedia.org/ontology/capital>
<http://dbpedia.org/resource/Country>

این خروجی بیانگر این است که رابطه Capital بین دو موجودیت City و Country برقرار می‌شود.

این عمل برای هریک از روابط سه‌تایی موجود در فایل نمونه‌ها تکرار می‌شود و بدین نحو نیاز به سرورهای قدرتمند و پردازش موازی را برای استفاده از این پایگاه شناخت بزرگ تا میزان چشم‌گیری کاهش می‌دهد.

۲-۷- تلخیص نمونه‌ها در پایگاه شناخت

یکی از ویژگی‌های بسیار پرکاربرد در چارچوب پیشنهادی، امکان خلاصه‌سازی پایگاه شناخت عمومی است. پایگاه شناخت‌های عمومی مثل DBpedia شامل موجودیت‌های بسیار زیاد هستند. حجم بالای داده موجب کند شدن بارگذاری پایگاه شناخت و هم‌چنین کند شدن موتور استنتاج می‌شود. لازم به ذکر است که بسیاری از داده‌های موجود در پایگاه شناخت عمومی به اخبار و موجودیت‌های فارسی ربطی ندارند. بسیاری دیگر نیز در حوزه خاصی که تمرکز سیستم پیش‌بینی بر آن است، نمی‌باشند. به همین خاطر خلاصه‌سازی پایگاه شناخت خدشه‌ای به عملکرد سیستم وارد نخواهد کرد. رویکرد ارائه شده برای تلخیص پایگاه شناخت شامل دو بخش است:

۲-۷-۱- خلاصه‌سازی دسته‌ها

کاربر دسته‌های مورد نظر خود را از لیست تمامی دسته‌های موجود در پایگاه شناخت عمومی (برای مثال DBpedia) انتخاب می‌کند. بعد از انتخاب دسته‌ها تمامی نمونه‌هایی که شامل دسته‌های مورد نظر هستند توسط سیستم انتخاب می‌شود. سپس کاربر می‌تواند مسیر مورد نظر را برای ذخیره پایگاه شناخت در قالب OWL ذخیره کند.

۲-۷-۲- خلاصه‌سازی ویژگی‌ها

این قسمت نیز بسیار شبیه به خلاصه‌سازی دسته‌ها است. با این تفاوت که کاربر از لیست کلی ویژگی‌ها تعدادی را انتخاب کرده و بعد از مشخص شدن نمونه‌هایی که توسط این ویژگی به هم متصل می‌شوند، ویژگی‌ها و نمونه‌ها در مسیر مورد نظر کاربر ذخیره می‌شود.

برای مثال اگر در حوزه‌ای که ما قصد پیش‌بینی رویدادهای آن را داریم مفاهیمی همچون شهر و کشور وجود داشته باشد، بایستی ضمن انتخاب دسته‌های Country، City و همچنین ویژگی Capital از پایگاه شناخت عمومی، در پایگاه شناخت خلاصه‌شده لحاظ شوند.

۳- پیش‌بینی رویدادهای اخبار

از آنجایی که مسأله پیش‌بینی رویدادهای اخبار مسأله بسیار جدیدی در حوزه علوم کامپیوتر است، تنها پرداختن به روش‌هایی که برای این مسأله خاص ارائه شده‌اند، نمی‌تواند مفید باشد. در واقع روش‌های بسیار کم و محدودی برای حل این مسأله ارائه شده است. به جز مطالعات صورت گرفته توسط رادینسکی [۲۳-۲۵]، ما از پژوهش‌های دیگر برای انجام فعالیتی که با آن مواجهیم، مطلع نیستیم: "دریافت رویدادهای دلخواه خبری در قالب زبان طبیعی و پیش‌بینی رویدادهایی که می‌توانند موجب آن‌ها شوند".

هدف ما در این پژوهش ارائه رویکردهایی برای انجام استنتاج علّی به منظور پیش‌بینی رویدادهای معلول در محیط‌های متنی است.

ما قصد توسعه یک روش جدید پیش‌بینی را داریم که با توجه به یک رویداد ارائه شده در قالب زبان طبیعی بتواند رویدادهای معلول آن را در آینده پیش‌بینی کند. برای این منظور، علاوه بر پیکره متنی اخبار (دانش پویا[1])، یک سری قواعد علّی خاص حوزه[2] از طریق یک واسط کاربری تعریف قواعد در قالب منطق مرتبه اول به عنوان دانش پیشین[3] یا دانش عام (بخشی از دانش ایستا[4]) به ماشین تزریق می‌شود. سپس از چندین پایگاه شناخت بزرگ شامل پایگاه شناخت‌های عمومی نظیر Wikipedia و DBpedia با بستر داده‌های پیوندی [۲۸]؛ پایگاه شناخت لغات نظیر WordNet [۲۹] و پایگاه شناخت افعال نظیر VerbNet [۳۰] به عنوان دانش ضمنی[5] (بخش دیگر دانش ایستا) برای تعمیم دادن رویدادها و تولید مدل پیش‌بینی استفاده می‌کند. این مدل در قالب استاندارد زبان پایگاه شناخت وب[6] (OWL) بازنمایی می‌شود که برای به دست آوردن مناسب‌ترین تعمیم‌های یک رویداد علت و پیش‌بینی رویدادهای معلول با استفاده از قواعد علی، استفاده می‌شود.

کار ما شامل چندین نوآوری است: اولین نوآوری ارائه یک روش جدید و مقیاسپذیر به منظور استفاده از تکنیک‌های معنایی و مستقل از زبان برای پیش‌بینی رویدادهای اخبار است. دوم اینکه ما یک روش جدید برای بازنمایی رویداد و همچنین تعریف و استفاده از قواعد علّی به عنوان بخشی از دانش عام در قالب منطق مرتبه اول برای پیش‌بینی رویدادهای جدید ارائه کرده‌ایم. سوم اینکه، ما چارچوبی برای ترکیب انواع مختلف دانش ایستا (به عنوان مثال دانش پیشین، دانش ضمنی) و همچنین تجمیع و ادغام آن با دانش پویا در قالب OWL پیشنهاد دادیم. در آخر، ما برای اولین بار برای آزمایش‌هایمان به منظور اندازه‌گیری صحت پیش‌بینی رویدادها از متون اخبار، یک روش خودکار و بدون دخالت انسان پیشنهاد کردیم که در کارهای قبلی مشاهده نشده است و این امکان را می‌دهد که علاوه بر دقت معیارهای غنی‌تری از عملکرد ماشین، نظیر پوشش و تنوع را نیز ارائه دهد.

۳-۱- چارچوب پیشنهادی

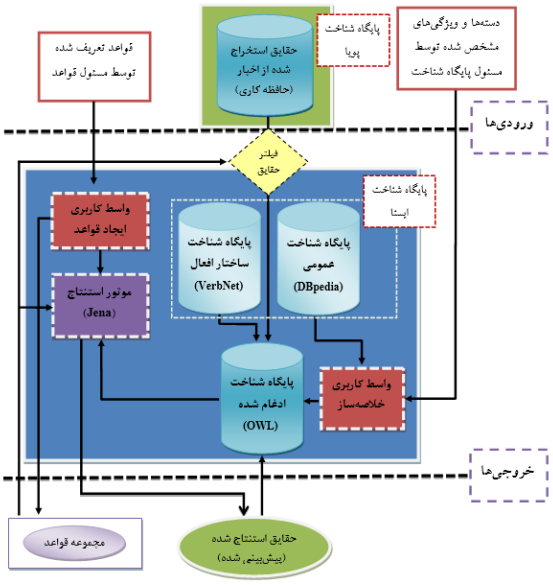
چارچوب ارائه شده با استفاده از دانش ضمنی به دست آمده در پایگاه شناخت و هم چنین آرگومان‌های به دست آمده از اخبار با ارضا کردن قواعدی که به ماشین تزریق شده است، می‌تواند رخدادهای جدید را پیش‌بینی کند. شکل ۱ زیربخش‌های اصلی چارچوب پیشنهادی را برای پیش‌بینی رویدادهای اخبار نشان می‌دهد.

در این شکل ورودی‌ها، خروجی‌ها و پردازش‌های انجام شده در بخش پیش‌بینی اخبار نشان داده شده است. دانش ضمنی از طریق یک پایگاه شناخت بزرگ (که در این پژوهش DBpedia فارسی شده است) به ماشین اضافه می‌شود. همان‌طور که در بخش ۲-۷ ذکر شد، یک واسط کاربری برای خلاصه‌سازی این پایگاه شناخت در چارچوب پیشنهادی در نظر گرفته شده است که با توجه به ورودی مسئول پایگاه شناخت (که دسته‌ها و ویژگی‌های مورد نیاز است) پایگاه شناخت ضمنی را خلاصه می‌کند. بعد از خلاصه‌سازی نتیجه در پایگاه شناخت اصلی سیستم ذخیره و بارگزاری می‌شود.

از طرف دیگر فعل‌ها و نقش‌های موضوعی به‌ترتیب در قالب دسته و ویژگی به پایگاه شناخت ساختار افعال اضافه می‌شوند. در این پایگاه شناخت هر فعل با یک دسته و هر نقش موضوعی به شکل یک ویژگی نمایش داده می‌شود. رویدادهای استخراج‌شده از اخبار (حافظه کاری[7]) نیز به‌صورت آرگومان‌های سه‌تایی (قالب معمول پایگاه شناخت) به سیستم اضافه می‌شود. این دو پایگاه شناخت نیز ضمن اضافه شدن به پایگاه شناخت اصلی با آن ادغام می‌شود. مسئول ایجاد قواعد، قاعده‌های مورد نیاز را در قالب پایگاه شناخت توسط واسط کاربری ایجاد کرده و در مسیر دلخواه در قالب پایگاه شناخت ذخیره می‌کند.

بعد از این مرحله کاربر می‌تواند از واسط کاربری استنتاج استفاده کند. توجه کنید که این واسط کاربری تنها رویدادهایی را از حافظه کاری دریافت می‌کند که با بخش اگر مربوط به قواعد تطابق داشته باشد. بدین نحو از پردازش‌های اضافی جلوگیری به عمل آمده و باعث تسریع عمل استنتاج می‌شود. نتایج استنتاج در قالب حقایق یا رویدادهایی خواهد بود که قابل تبدیل به آرگومان‌های سه‌تایی (در قالب OWL) است. کاربر می‌تواند بعد از استنتاج این نتایج را در قالب فایل‌های پشتیبانی شده پایگاه شناخت به پایگاه شناخت اصلی اضافه و ادغام کند. در ادامه به شرح قابلیت‌های موجود در چارچوب پیشنهادی پرداخته می‌شود.

[1] Dynamic knowledge
[2] Domain specific
[3] A priori knowledge
[4] Static knowledge
[5] Contextual knowledge
[6] Ontology Web Language
[7] Working Memory



شکل ۱: چارچوب پیشنهادی برای پیش بینی رویدادهای اخبار.

۳-۲- افزودن حقایق موجود در متن اخبار در قالب OWL

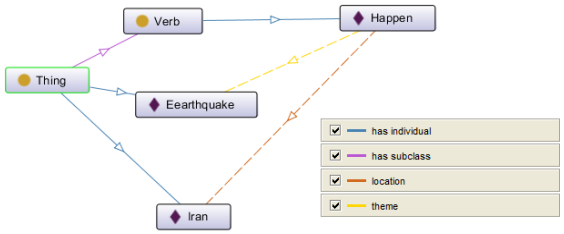
سیستم پیش‌بینی با توجه به رویدادهای اخیر و در نظر گرفتن قواعد علّی می‌تواند رویدادهای آتی را پیش‌بینی کند. مفاهیم و رویدادهای گذشته باید در قالب پایگاه شناخت قرار بگیرد تا برای پیش‌بینی استفاده شود. به همین منظور هر جمله به شکل چندین آرگومان در پایگاه شناخت ذخیره می‌شود. در مدل ارائه شده برای بیان هر جمله به صورت آرگومان‌هایی در پایگاه شناخت از روش فعل-محور[1] استفاده می‌شود. به این معنی که هسته تمام حقایق استخراج شده از اخبار، فعل جمله و یا جملات خبر است. مرکز هر جمله، فعل جمله در نظر گرفته شده است و به ازای هر جمله خبری یک نمونه از فعل مربوط به آن جمله در پایگاه شناخت قرار می‌گیرد.

در این مدل هر جمله (حقیقت) شامل یک فعل و تعدادی نقش موضوعی است که به فعل متصل می‌شوند. به‌طوری که در حین افزودن جملات (حقایق) به پایگاه شناخت هر فعل با یک دسته و هر نقش موضوعی به شکل یک ویژگی نمایش داده می‌شود. در واقع هر نقش موضوعی یک فعل را به نمونه‌ای از کلاس‌های دیگر متصل می‌کند. برای نمونه جمله زیر را در نظر بگیرید:

An earthquake happened in Iran.

گراف ترسیم شده در شکل (۲) نحوه نمایش جمله را در پایگاه شناخت نشان می‌دهد:

Verb: happen
Theme: Earthquake
Location: Iran



شکل ۲: نحوه نمایش جمله در پایگاه شناخت.

توجه کنید که Happen یک نمونه است. در پایگاه شناخت می‌توانیم بیشمار نمونه از این نوع فعل داشته باشیم. برای مثال *happen_001*، که عدد مربوط به این فعل (یعنی 001) می‌تواند به عنوان شناسه خبر در نظر گرفته شود. لذا حقایق استخراج شده برای این جمله شامل سه آرگومان به شکل زیر خواهد بود:

<happen_001> < type> < happen>.
<happen_001> < theme> < Earthquake>.
<happen_001> < location> < Iran>.

بخش استخراج حقایق، نقش‌های موضوعی هر جمله را نیز استخراج می‌کند. تعداد و نوع این نقش‌ها از پیش تعیین شده است. برای این منظور از مدل کِیم[2] [۳۱] برای بازنمایی رویدادها و حقایق بهره گرفته شد. در این مدل فرض می‌شود مجموعه مشخصی از موجودیت‌ها در دسترس است. برای مثال این موجودیت‌ها می‌توانند اشیا، انسان‌ها یا مفاهیم مجرد باشند. در این روش رویدادها توسط یک سه‌تایی مرتب به شکل $[O, P, T]$ نشان داده می‌شود. در این سه‌تایی O مجموعه از موجودیت‌ها، P یک رابطه یا ویژگی است و T یک بازه زمانی را نشان می‌دهد.

در مدل ارائه شده برای بازنمایی پیش‌بینی رویداد، مدل کِیم [۳۱] برای بازنمایی رویدادها تعمیم داده شده است. به این ترتیب که هر رویداد شامل یک کنش یا فعل (P) است؛ در مجموعه موجودیت‌ها (O)، عملگر کنش یا اجزای مورد بحث در رویداد قرار می‌گیرند و در انتها نیز زمان وقوع کنش (T) قرار می‌گیرد. در این شیوه، بازنمایی هر رویداد به شکل $<P, O_1, O_2,\ldots, O_{13}, T>$ پیشنهاد می‌شود، که در آن متغیرهای O_1 تا O_7 شامل برچسب‌های کنش و متغیرهای O_8 تا O_{13} حاوی برچسب‌های توصیف در نقش‌های موضوعی می‌باشند. در جدول (۲)، لیست نقش‌های موضوعی

[1] Verb-based

[2] Kim

پیشنهادی برای هر متغیر در مدل تعمیم‌یافته کِیم آورده شده است.

برای مثال رویداد "تولید اورانیم توسط ایران با غنای حداقل ۶۰ درصد به شیوه جدید با استفاده از سانتریفیوژ در اراک در بهمن‌ماه ۹۳" را درنظر بگیرید. متغیرهای مربوط به مدل تعمیم‌یافته در رویداد مذکور به شکل تولید=P، ایران=O_1، اورانیم=O_3، اراک=O_4، شیوه جدید=O_5، استفاده از سانتریفیوژ=O_6، غنای=O_8، حداقل=O_{12}، ۶۰ درصد=O_{13} و بهمن‌ماه ۹۳=T بازنمایی می‌شود. توجه شود برای متغیرهایی نظیر O_2، O_7، O_9 و... که در جمله نقش معادلی ندارند، مقدار تهی (Null) درنظر گرفته می‌شود.

جدول ۲: جایگاه نقش‌های معنایی پیشنهادی در متغیرهای مدل تعمیم‌یافته کِیم.

توصیف نقش	نقش معنایی در مدل تعمیم‌یافته
فعل انجام کنش	P = Action
عامل انجام کنش	O_1 = Agent
تاثیرپذیرفته در کنش	O_2 = Patient
موضوع مطرح در کنش	O_3 = Theme
مکان انجام کنش	O_4 = Location
شیوه انجام کنش	O_5 = Manner
ابزار انجام کنش	O_6 = Instrument
علت انجام کنش	O_7 = Cause
توصیف انجام کنش	O_8 = Description
بخشی از (متعلق به)	O_9 = Part-of (Has)
نوعی از	O_{10} = Is-a
نمونه‌ای از	O_{11} = Instanse-of
مقدار/ تعداد	O_{12} = Value
واحد	O_{13} = Measurement
زمان وقوع کنش	T = Time

۳-۳- افزودن قواعد کاربر در قالب منطق مرتبه اول

همان‌طور که گفته شد هر قاعده شامل تعدادی گره است. این گره‌ها می‌تواند در بخش *اگر* یا بخش *آنگاه* در قاعده اضافه شود. در واقع با ارضا شدن تمام گره‌های بخش *اگر* گره‌های بخش *آنگاه* ارضا می‌شود. گره‌های هر دو بخش یکی از ۴ نوع از پیش تعیین شده هستند. این ۴ نوع شامل *دسته*، *ویژگی*، *فعل* و *نقش موضوعی* است. در ادامه، با ذکر یک مثال نمونه‌ای از ایجاد قواعد بیان شده است.

فرض کنید می‌خواهیم قاعده‌ای به مجموعه قواعد اضافه کنیم که *«چنانچه در کشوری زلزله اتفاق بیفتد در پایتخت آن کشور خرابی ایجاد می‌شود»*. اگرچه این قاعده خیلی منطقی نیست ولی برای نشان دادن روند تولید قواعد مناسب است چرا که تمام حالت‌های گره‌ها در این قواعد لحاظ شده است.

قاعده مورد نظر به شکل زیر به مجموعه قواعد اضافه خواهد شد:

```
[rule1:
(?A theme earthquake)          نقش موضوعی
,(?A rdf:type happen)          فعل
,( ?A location ?B)             نقش موضوعی
,( ?B capital ?C)              ویژگی
->
(?D rdf:type destroy)          فعل
,(?D location ?C)]             نقش موضوعی
```

هر کدام از خطوط قاعده نوشته شده یک گره است. همان‌طور که ذکر شد هر گره ۴ نوع گزاره را می‌تواند بیان کند. روبروی هر گره نوع گزاره آن آورده شده است. برای مثال فرض کنید کاربر می‌خواهد گره اول یعنی (*?A* theme earthquake) را تولید کرده و به قسمت *اگر* قاعده اضافه کند. ابتدا کاربر باید ضمن انتخاب نقش موضوعی theme برای نوع گره و متغیر *?A* برای آن مقدار earthquake را به عنوان مقدار برای این نقش موضوعی وارد کند. به طور مشابه این فرآیند می‌تواند برای اضافه کردن سایر گره‌ها به قاعده موردنظر تکرار شود.

بعد از تولید تمامی قواعد دلخواه، مجموعه قواعد در یک فایل با پسوند rule. در قالب قابل استفاده در Jena ذخیره می‌شود. این فایل در قسمت استنتاج مورد استفاده قرار می‌گیرد.

۴-۳- استنتاج در OWL در قالب منطق مرتبه اول

در این بخش کاربر می‌تواند پایگاه شناخت مربوط به دانش عمومی (خلاصه شده DBpedia) و پایگاه شناخت مربوط به اخبار را در سیستم بارگزاری کند و سپس با اضافه کردن مجموعه قواعد ذخیره شده حقایق جدیدی که قابل استنتاج است را بازیابی نماید. بعد از انجام استنتاج، نتایج که شامل لیستی از آرگومان‌های جدید (به صورت سه‌تایی‌هایی در قالب پایگاه شناخت) است به کاربر نمایش داده شده و در صورت تأیید کاربر در پایگاه شناخت اصلی ذخیره می‌شود.

برای مثال فرض کنید خبری تحت عنوان زیر منتشر می‌شود:

An earthquake happened in Iran.

با توجه به مطالب بیان‌شده در بخش ۲-۵، حقایق استخراج شده شامل سه آرگومان به شکل زیر خواهد بود:

<happen_001> < type> < happen>.
<happen_001> < theme> < Earthquake>.
<happen_001> < location> < Iran>.

توجه کنید همان‌طور که در بخش ۲-۷ نیز توضیح داده شد، بایستی ابتدا دسته‌های Country، City و ویژگی Capital از پایگاه شناخت عمومی از پیش انتخاب شده باشند.

قابل ذکر است که تمام نمونه‌ها و دسته‌ها در پایگاه شناخت به صورت آدرس اینترنتی URL ذخیره می‌شود. برای سهولت کاربر در بخش ایجاد قوانین، خلاصه‌سازی پایگاه شناخت و استنتاج آدرس کامل اینترنتی نمایش داده نمی‌شود ولی در فایل‌های ذخیره شده به شکل کامل ذخیره خواهد شد.

نتیجه به دست آمده از استنتاج برای این مثال شامل دو آرگومان به شکل زیر خواهد بود:

<destroy_001> < type> < Destroy>.
< destroy_001> < location> < Tehran>.

آرگومان اول نشان می‌دهد که این نمونه از دسته فعل Destroy است. آرگومان دوم این نمونه را با نقش موضوعی location به تهران متصل می‌کند. در واقع اگر بخواهیم این دو آرگومان را به شکل یک جمله نمایش دهیم باید بگوییم:

Destruction in Tehran.

۴- ارزیابی پیش‌بینی

۴-۱- دادگان موردنیاز ارزیابی

به طور کلی، دادگان مورد نیاز برای ارزیابی شامل پایگاه شناخت ایستا (دانش استخراج شده از DBpedia)، پایگاه شناخت پویا (حقایق استخراج شده از اخبار) و پایگاه قواعد مرتبط با اخبار (قواعد استخراج شده از افراد خبره) است. در بخش ارزیابی، متون اخبار انتخابی و قواعد استنتاج اهمیت بسیاری دارد چرا که قدرت استنتاج وابسته به این دو است. هم‌چنین ارزیابی سیستم تنها با داده‌ها امکان پذیر نیست و باید سیستم توسط انسان ارزیابی شود که نحوه انجام این کار در بخش ۴-۳ توضیح داده شده است.

۴-۲- معیارهای ارزیابی

از آنجایی که تا کنون روشی برای حل مسأله پیش‌بینی رویدادهای آینده در متون اخبار فارسی به این شکل ارائه نشده است، برای ارزیابی روش پیشنهادی، عملکرد آن در مقایسه با عملکرد انسان در پیش‌بینی رویداد مورد بررسی قرار گرفته است. در این آزمایش از انسان‌ها و ماشین پرسیده می‌شود که چه رویدادهایی با مشاهده رویداد آزمون ممکن است در آینده اتفاق بیفتد. یکی از شاخص‌هایی که برای ارزیابی پاسخ‌ها در پژوهش‌های مرتبط مورد بررسی قرار گرفته است، مقایسه میزان صحت[1] پیش‌بینی‌های تولید شده توسط انسان و ماشین است. طبق تعریف این شاخص درصد پیش‌بینی‌هایی است که در آینده اتفاق افتاده است. اما کار با این شاخص دو مشکل در پی خواهد داشت. اول اینکه ممکن است پیش‌بینی مورد نظر بسیار محتمل باشد ولی با این حال در دنیای واقعی اتفاق نیفتد. مشکل دوم این است که شاید پیش‌بینی درست بوده باشد ولی علت رویداد آن آزمون نباشد و در واقع مسائل دیگری در دنیای واقعی در وقوع آن دخیل باشند. به همین خاطر برای بررسی و مقایسه روش پیشنهادی در برخی از مطالعات [۲۳] معیار دیگری با نام کیفیت[2] نیز تعریف شده است. این معیار احتمال وقوع رویداد پیش‌بینی شده را بر اساس رویداد آزمون نشان می‌دهد.

همان‌طور که اشاره شد، بیشتر روش‌های ارزیابی در کارهای قبلی تنها بر روی صحت و خطای پیش‌بینی‌های تولید شده براساس ارزیاب انسانی متمرکز شده است. این درحالیست که افراد بیشتر تمایل دارند لیستی از رویدادهای موردعلاقه آنها پیش‌بینی و به آنها ارائه شود. از این‌رو، معیارهای صحت و خطا (نظیر میانگین خطای مطلق[3]، میانگین خطای مربع[4] و ریشه میانگین خطای مربع[5]) دارای کارایی لازم نمی‌باشند. لذا ما در اینجا از معیار دقت[6] برای ارزیابی روش پیشنهادی و مقایسه پیش‌بینی‌های تولید شده توسط آن با سایر روش‌ها بهره می‌گیریم (این معیار در واقع نرمال شده معیار کیفیت است). اما این معیار نیز به تنهایی نشان‌دهنده‌ی عملکرد بهتر یک روش پیش‌بینی نیست. زیرا کاربران تمایل دارند پیش‌بینی‌های تولید شده علاوه بر دقت از پوشش[7] و تنوع[8] بالایی نیز برخوردار باشند. لذا در اینجا از معیارهای دقت، پوشش و تنوع برای ارزیابی روش پیشنهادی استفاده می‌کنیم.

از روابط (۱) و (۲) بترتیب برای محاسبه دقت و پوشش سیستم پیش‌بینی اخبار استفاده می‌شود.

$$Precision = \frac{|\{Extracted\ events\} \cap \{Inferred\ events\}|}{|\{Inferred\ events\}|}, \qquad (۱)$$

[1] Accuracy
[2] Quality
[3] Mean Absolute Error (MAE)
[4] Mean Square Error (MSE)
[5] Root Mean Square Error (RMSE)
[6] Precision
[7] Coverage
[8] Diversity

$$Coverage = \frac{|\{Extracted\ events\} \cap \{Inferred\ events\}|}{|\{Extracted\ events\}|}, \quad (۲)$$

تنوع بین‌لیستی[1] پیش‌بینی‌های تولید شده برای دو حوزه خاص اخبار از طریق محاسبه‌ی فاصله همینگ[2] [۳۲] بین لیست‌های پیش‌بینی شده برای هریک بدست می‌آید. فاصله همینگ ((H(L)) با محاسبه اختلاف هر عنصر از یک لیست و عنصر متناظر آن در لیست دیگر و جمع قدر مطلق اختلاف‌ها به دست می‌آید. رابطه (۳)، نحوه محاسبه تنوع بین دو لیست i و j به طول $|L|$ را نشان می‌دهد. در این رابطه، $|Q_{ij}(L)|$ برابر با تعداد رویدادهای یکسان و مشترک پیش‌بینی شده به صورت متناظر در دو لیست i و j می‌باشد.

$$D_{Inter} = H(L) - 1 - \frac{|Q_{ij}(L)|}{|L|}, \quad (۳)$$

بر این اساس، تنوع درون‌لیستی[3] پیش‌بینی‌های تولید شده برای دو حوزه اخبار را نیز به صورت نشان داده شده در رابطه (۴) تعریف می‌کنیم.

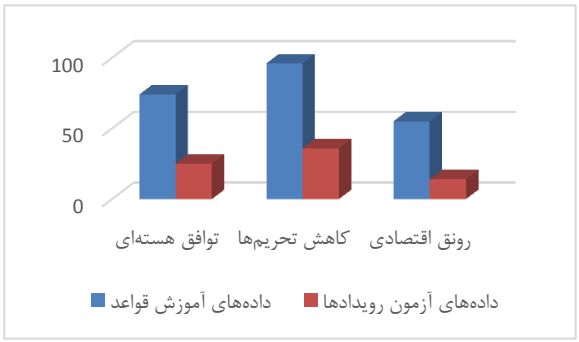
$$D_{Intra} = 1 - H(L). \quad (۴)$$

۴-۳- رویه‌ی ارزیابی

بخش ارزیابی سیستم پیش‌بینی بسیار پیچیده خواهد بود چرا که داده‌های آموزشی با برچسب مشخص در دسترس نیست. یعنی برای هر خبر رویداد خاصی به عنوان خروجی پیش‌بینی مدنظر نداریم. در واقع همانند دیگر مسائل هوش مصنوعی با فقدان داده‌های آموزشی همراه با برچسب‌های صحیح قطعی روبرو هستیم. به همین منظور در بخش ارزیابی دو گروه از افراد مطلع به اخبار داخلی و سیاست خارجی برای تولید داده‌ها جهت ارزیابی ماشین در نظر گرفته شدند. مجموعه‌ای از متون اخبار در حوزه خاص برای مثال مسأله هسته‌ای ایران انتخاب شده و در اختیار گروه اول قرار می‌گیرد. این گروه باید تعدادی نمونه رویداد علت و معلول را از اخبار استخراج کنند. تا حد امکان بایستی این افراد متن اخبار را مورد توجه قرار داده و از دخالت دیگر اخبار در ذهن خود در این نمونه‌ها خودداری کنند. قابل توجه است که رویدادهای علت استخراج شده توسط افراد بایستی لزوماً در متون اخبار آزمون موجود باشد، ولی برای استخراج رویدادهای معلول، افراد می‌توانند علاوه بر این اخبار از سایر اخبار منتشر شده در حوزه خاص نیز استفاده نمایند.

گروه دوم مسئول ایجاد قواعد در قالب تعریف شده (به صورت منطق مرتبه اول) با توجه به متون اخبار هستند. چون این گروه از نمونه‌های انتخاب شده توسط گروه اول بی‌خبر هستند نظر آن‌ها دقیقاً منطبق بر مرحله قبل نخواهد بود. این گروه نیز باید نهایت توجه را در استخراج قواعد پیش‌بینی با توجه به متون اخبار مبذول دارند. برای این منظور تنها رویدادهای علت استخراج شده توسط گروه اول در اختیار این گروه قرار گرفته و افراد این گروه – که از رویدادهای معلول استخراج شده توسط گروه اول بی‌خبر هستند – ضمن انتخاب رویدادهای معلول جدید، قواعدی را با توجه به این نمونه‌های جدید استخراج می‌کنند.

برای ارزیابی سیستم از مجموعه داده‌های مشتمل بر ۲۰۰ جفت رویداد استخراج شده به صورت تصادفی از اخبار جمع‌آوری شده در حوزه سیاست خارجی شامل موضوعات توافق هسته‌ای (D1)، کاهش تحریم‌ها (D2) و توسعه رونق اقتصادی (D3) استفاده شده است. نمودار شکل ۳، توزیع داده‌های آموزش و آزمون را برای ارزیابی سیستم پیشنهادی نشان می‌دهد.



شکل ۳: نمای توزیع داده‌های جمع‌آوری شده برای چندین حوزه اخبار سیاست خارجی.

در این مجموعه داده، تمامی رویدادهای علت و معلول آموزش جهت تعمیم و تولید قواعد در اختیار افراد خبره قرار می‌گیرد. سپس، رویدادهای علت آزمون همگی به سیستم داده می‌شود و رویداد معلول توسط سیستم پیش‌بینی تولید می‌شود. نتایج تولید شده به طور خودکار با نتایج واقعی مقایسه شده و دقت، پوشش و تنوع پیش‌بینی‌های تولید شده ارزیابی می‌گردد.

۴-۴- نتایج ارزیابی

برای اینکه با نتایج پیش‌بینی سیستم پیشنهادی با استفاده از پایگاه شناخت فارسی استخراج شده تا حدی آشنا شوید، در اینجا مثالی از این نوع پیش‌بینی ارائه می‌شود. فرض کنید، قاعده‌ای داریم که «*چنانچه در کشوری زلزله اتفاق بیفتد در پایتخت آن*

[1] Inter-diversity
[2] Hamming distance
[3] Intra-diversity

کشور خرابی ایجاد می‌شود». حال فرض کنید جمله جدیدی تحت عنوان «*رخداد زلزله در ایران*» منتشر می‌شود.

موتور پیش‌بینی بایستی پس از استخراج معنایی حقایق از متن ورودی جمله و قرار دادن آن در قالب پایگاه شناخت و همچنین با اهتمام به دانش‌های ضمنی «*ایران یک کشور است*» و «*تهران پایتخت ایران است*» - که از منبع پایگاه شناخت فارسی استخراج شده از DBpedia استفاده شده است - قادر به استنتاج جمله جدیدی تحت عنوان «*خرابی در تهران*» باشد.

در جدول (۳)، نمونه‌ای از استنتاج‌های تولید شده توسط ماشین برای پیش‌بینی رویدادهای استخراج شده از اخبار آورده شده است.

جدول ۳: نمونه‌ای از پیش‌بینی‌های استنتاج شده برای رویدادهای استخراج شده از اخبار.

رویداد استنتاج شده (معلول ۲)	رویداد استنتاج شده (معلول ۱)	رویداد استخراج شده (علت)
افزایش آرامش بین المللی	برداشته شدن تحریم‌ها علیه ایران	وارد شدن مذاکره ایران با گروه ۵+۱
کاهش نرخ طلا	رونق مجدد مناسبات اقتصادی	حصول توافق هسته‌ای میان ایران و گروه ۵+۱
پایین آمدن سطح معیشت مردم	تولید و توسعه فساد	ادامه اعلام تحریم اقتصادی
تسریع دستیابی ایران به فناوری هسته‌ای	افزایش اعتماد آژانس بین‌المللی انرژی اتمی	تعلیق غنی‌سازی ۲۰ درصد به صورت داوطلبانه

همان‌طور که در [۳۳] اشاره شده است، هیچ مطالعه‌ای که بتوان این وظیفه را با آن ارزیابی کرد، وجود ندارد و معمول‌ترین و هزینه‌برترین رویکرد ارزیابی توسط انسان می‌باشد. ما نیز از هیچ فعالیتی که بتوان این وظیفه را به صورت پیشنهاد شده در این پژوهش با آن ارزیابی کرد، مطلع نیستیم. از این‌رو، ما نتایج روش پیشنهادی را با دو روش پایه مقایسه می‌کنیم: (۱) استنتاج در منطق مرتبه اول بدون استفاده از پایگاه شناخت؛ (۲) استفاده از تخمین افراد خبره برای انجام پیش‌بینی این رویدادها.

برای این منظور، پیش‌بینی رویدادها در سه نمونه پرس‌وجوی خاص اخبار جمع‌آوری شده در حوزه سیاست خارجی شامل D1، D2 و D3 صورت گرفته است. نتایج تجربی مقایسه میزان دقت، پوشش و تنوع درون‌لیستی پیش‌بینی اخبار برای روش‌های مختلف به‌ترتیب در جدول‌های (۴)، (۵) و (۶) نشان داده شده است.

جدول ۴: مقایسه میزان دقت پیش‌بینی برای چندین حوزه از اخبار سیاست خارجی.

Method	Nuclear deal	Sanctions relief	Development of economic relations
Proposed (News + Ontology)	86.74%	79.65%	84.56%
News source alone	46.68%	39.08%	41.23%
Human expert	62.37%	58.94%	71.63%

جدول ۵: مقایسه میزان پوشش اخبار پیش‌بینی شده برای چندین حوزه سیاست خارجی.

Method	Nuclear deal	Sanctions relief	Development of economic relations
Proposed (News + Ontology)	88.13%	81.36%	79.34%
News source alone	64.47%	61.38%	57.89%
Human expert	56.73%	51.29%	47.03%

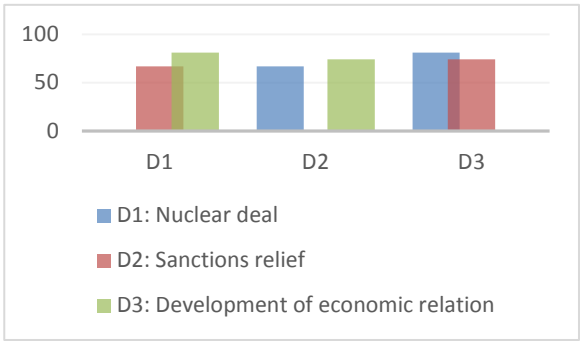
جدول (۶)، مقایسه میزان تنوع درون‌لیستی اخبار پیش‌بینی شده را با استفاده از روش‌های مختلف و برای حوزه‌های D1، D2 و D3 نشان می‌دهد.

جدول ۶: مقایسه تنوع درون‌لیستی اخبار پیش‌بینی شده برای چندین حوزه سیاست خارجی.

Method	Nuclear deal	Sanctions relief	Development of economic relations
Proposed (News + Ontology)	83.46%	74.39%	79.78%
News source alone	42.67%	36.13%	38.41%
Human expert	59.14%	55.29%	66.58%

همان‌طور که در جدول‌های (۴) تا (۶) قابل مشاهده است، عملکرد روش پیشنهادی در تمام موارد بهتر از روش پایه می‌باشد. علت برتری روش پیشنهادی در مقایسه با استنتاج در منطق مرتبه اول بدون استفاده از پایگاه شناخت، به‌طور واضح به تعمیم نقش‌های موضوعی استخراج شده با استفاده از پایگاه شناخت برمی‌گردد. همان‌طور که نتایج تجربی نشان داد، استفاده از تخمین افراد خبره برای انجام پیش‌بینی رویدادها در مقایسه با ماشین، که با توجه به گستره‌ی ویژگی‌ها و مفاهیم موجود در متون اخبار و همچنین نیاز به بررسی و تعامل با حجم بسیار زیادی از داده‌ها، حجم وسیعی از اخبار گذشته را همراه با روابط علّی بین رویدادها با استفاده از پایگاه‌های شناخت موجود در شبکه جهانی اینترنت بررسی می‌کند، عملکرد کمتری داشته است. علاوه بر این استفاده از مزایای منطق برای بازنمایی رویدادها نیز به عنوان راه‌حلی برای پیچیدگی مسئله، بر بهبود این عملکرد برای ماشین بی‌تاثیر نبوده است.

در نمودار شکل (۴)، میزان تنوع بین‌لیستی اخبار پیش‌بینی شده با استفاده از روش پیشنهادی برای حوزه‌های مختلف سیاست خارجی نمایش داده شده است. همان‌طور که قابل مشاهده است، نتایج پیش‌بینی اخبار برای حوزه‌ی D3 دارای بیشترین تنوع و برای حوزه‌ی D2 دارای کمترین تنوع نسبت به سایر حوزه‌های مشابه می‌باشد.



شکل ۴: تنوع بین‌لیستی اخبار پیش‌بینی شده برای چندین حوزه سیاست خارجی.

علاوه بر این، ما آزمایش دیگری نیز به منظور بررسی تطابق معنایی هر یک از نقش‌های موضوعی با پایگاه شناخت ایجاد شده در زبان فارسی ترتیب دادیم. هدف از این آزمایش نشان دادن این موضوع است که آیا نقش‌های استخراج شده از متون اخبار، از لحاظ معنایی، به درستی به یک مفهوم در پایگاه شناخت نگاشت شده‌اند. نتایج این آزمایش به‌طور خلاصه در جدول (۷) آمده است. همان‌طور که قابل مشاهده است، صحت تطابق برای فعل‌ها، بالاتر از سایر نقش‌های معنایی استخراج شده از اخبار است. این امر، به علت استفاده از پایگاه شناخت تخصصی Verbnet (استخراج شده از Farsnet) علاوه بر پایگاه شناخت فارسی (استخراج شده از DBpedia) می‌باشد.

۵- نتیجه‌گیری و پیشنهاد برای پژوهش‌های آتی

همان‌طور که بیان شد مسأله پیش‌بینی رویداد یکی از مسائل جدید در حوزه علوم کامپیوتر و هوش مصنوعی است. نیازهای اصلی این مسأله که در واقع پردازش زبان طبیعی با دقت مطلوب است، وظیفه‌ای دشوار برای ماشین است. در بخش پژوهش‌های پیشین تعدادی از رویکردهای یادگیری ماشین به این مسأله بیان شد. از مزایای این روش‌ها در پیش‌بینی رویدادها می‌توان به موارد زیر اشاره کرد: ۱) استفاده از حجم عظیم داده‌های متنی برای تولید جفت رویدادها بعنوان داده‌های آموزشی؛ ۲) تعمیم در سطح افعال و موجودیت‌های رویدادها. همچنین معایب این روش‌ها نیز عبارتند از: ۱) نیاز به سیستم پردازش زبان طبیعی با دقت بالا؛ ۲) رویکرد نامناسب برای استخراج جفت رویدادها با پوشش پایین؛ و ۳) پردازش بسیار پیچیده در بخش تعمیم و اجرای الگوریتم استنتاج.

جدول ۷: بررسی میزان صحت تطابق نقش‌های معنایی استخراج شده از اخبار با مفاهیم موجود در پایگاه شناخت.

نقش موضوعی	توصیف نقش	صحت تطابق در پایگاه شناخت
P = Action	فعل انجام کنش	93%
O_1 = Agent	عامل انجام کنش	89%
O_2 = Patient	تاثیرپذیرفته در کنش	84%
O_3 = Theme	موضوع مطرح در کنش	78%
O_4 = Location	مکان انجام کنش	85%
O_5 = Manner	شیوه انجام کنش	65%
O_6 = Instrument	ابزار انجام کنش	74%
O_7 = Cause	علت انجام کنش	69%
O_8 = Description	توصیف انجام کنش	46%
O_9 = Part-of (Has)	بخشی از (متعلق به)	54%
O_{10} = Is-a	نوعی از	56%
O_{11} = Instanse-of	نمونه‌ای از	39%
O_{12} = Value	مقدار/ تعداد	41%
O_{13} = Measurement	واحد	46%

با توجه به مشکلات مذکور، سیستمی مبتنی بر قاعده پیشنهاد دادیم به شکلی که دانش عمومی و دانش موجود در اخبار را در پایگاه شناخت (آنتولوژی) ذخیره کرده و با چارچوب تعریف شده برای قواعد و با تأکید بر رفع نیازمندی‌های زیر قادر به استنتاج مورد نظر کاربر باشد: (۱) بازنمایی رویدادها در قالب نقش‌های معنایی در OWL؛ (۲) استفاده از پایگاه دانش عمومی جهت تعمیم رویدادها؛ (۳) تجمیع دانش ایستا و دانش پویا در یک پایگاه شناخت واحد؛ و (۴) تعریف قواعد در قالب FOL برای استنتاج.

به علت گسترده بودن موضوع پژوهش، زمینه‌های موجود برای فعالیت‌های آتی متنوع خواهد بود. یکی از مهم‌ترین این زمینه‌ها سعی در ایجاد الگوریتم خودکار یادگیری ماشین برای استخراج روابط علّی در خبرها است. در پژوهش [۱۵] روابط علّی تنها از عنوان اخبار استخراج شده و بسیاری از این روابط نیز توسط الگوریتم تشخیص داده نشده و چشم‌پوشی می‌شوند. برای این منظور می‌توان الگوریتم‌های کارایی همراه با بازخوانی بالاتر ارائه داد.

از طرف دیگر با وجود مزایای بسیار زیاد منطق که قادر به پوشش

حجم وسیعی از دانش بشر است، استفاده از این ابزار به تنهایی نمی‌تواند در محیط‌های غیرقطعی که مسأله ما نیز با آن مواجه است، مفید واقع گردد. از این‌رو ترکیب مدل‌های احتمالاتی و منطق می‌تواند به‌عنوان یکی از مهم‌ترین پژوهش‌های آتی – که در حال حاضر توسط نویسندگان این مقاله در دست مطالعه است - در نظر گرفته شود.

بخش بسیار مهم دیگر در این فعالیت استفاده از دانش عمومی موجود در وب است. همان‌طور که ذکر شد در سیستم فعلی از پایگاه شناخت عمومی DBPedia استفاده شده است. ترجمه این پایگاه شناخت و ساخت معادل کارایی از این پایگاه شناخت برای زبان فارسی به عملکرد این سیستم بسیار کمک خواهد کرد. البته روش محدودی در این پژوهش ارائه شده است، که توسعه آن در فازهای آتی پژوهش مطرح خواهد بود.

سپاسگزاری

این پژوهش تحت حمایت و بودجه پروژه تحلیل اخبار با مدیریت دکتر مریم حورعلی، دانشگاه صنعتی مالک اشتر و مشاوره دکتر هشام فیلی در دانشگاه تهران به انجام رسیده است. بدین وسیله نویسندگان مقاله از حمایت‌های بی‌دریغ ایشان کمال سپاسگزاری را دارند.

مراجع

[1] T. Kahneman, "On the Psycology of Prediction," Psycological Review, vol. 80, pp. 237-251, 1973.

[2] M. Banko, M. J. Cafarella, S. Soderl, M. Broadhead, and O. Etzioni, "Open information extraction from the web," in *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*, 2007.

[3] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka, and T. Mitchell, "Toward an architecture for never-ending language learning," in *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2010.

[4] C. Khoo, S. Chan, and Y. Niu, "Extracting causal knowledge from a medical database using graphical patterns," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2000, pp. 336-343.

[5] R. Girju and D. Moldovan, "Text mining for causal relations," in *Proceedings of the Annual International Conference of the Florida Artificial Intelligence Research Society (FLAIRS)*, 2002, pp. 360-364.

[6] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web", *Communications of the ACM - Surviving the data deluge*. 2008, pp. 68-74.

[7] D. B. Lenat and R. V. Guha, Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[8] M. Palmer, D. Gildea, and N. Xue, "Semantic role labeling," Synthesis Lectures on Human Language Technologies, vol. 3, pp. 1-103, 2010.

[9] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL)*, Hong Kong, 2000, pp. 512-520.

[10] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz, "Modeling and predicting behavioral dynamics on the web," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 599-608.

[11] G. Amodeo, R. Blanco, and U. Brefeld, "Hybrid models for future event prediction," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, 2011, pp. 1981-1984.

[12] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proceedings of the 2010 IEEE/ WIC/ ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2010.

[13] R. P. Schumaker and H. Chen, "extual analysis of stock market prediction using breaking financial news: The AZFin text system," ACM Transactions on Information Systems (TOIS), vol. 27, p. 12, 2009.

[14] P. Falinouss, "Stock trend prediction using news articles: a text mining approach," Master Thesis, Department of Business Administration and Social Sciencees, Luleå University of Technology, 2007.

[15] K.-Y. Chen, L. R. Fine, and B. A. Huberman, "Predicting the Future," *Information Systems Frontiers*, vol. 5, pp. 47–61, 2003.

[16] D. M. Pennock, S. Lawrence, C. L. Giles, and F. A. Nielsen, "The real power of artificial markets," *Science*, vol. 291, pp. 987-988, 2001.

[17] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the American Society for Information Science and Technology*, vol. 60, pp. 2169-2188, 2009.

[18] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, vol. 1, pp. 1-46, 2007.

[19] I. Zukerman, D. W. Albrecht, and A. E. Nicholson, "Predicting users' requests on the WWW," presented at *The 7th International Conference on User Modeling (UM99)*, Banff, Canada, 1999.

[20] K. Radinsky, S. Davidovich, and S. Markovitch, "Predicting the news of tomorrow using patterns in web search queries," in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/ WIC/ ACM International Conference on*, 2008, pp. 363-367.

[21] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012-1014, 2009.

[22] J. Durbin and S. Koopman, "Time Series Analysis by State Space Methods," *Oxford University Press*, 2008.

[23] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning to Predict from Textual Data," *Journal of Artificial Intelligence Research*, vol. 45, pp. 641-684, 2012.

[24] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning causality for news events prediction," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 909-918.

[25] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning causality from textual data," in *Proceedings of Learning by Reading for Intelligent Question Answering Conference*, 2011.

[26] K. Chan and W. Lam, "Extracting causation knowledge from natural language texts," *International Journal of Information Security (IJIS)*, vol. 20, pp. 327-358, 2005.

[27] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Kleef, S. Auer, and C. Bizer, "DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia," *Semantic Web*, vol. 1, pp. 1-5, 2012.

[28] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, 2009, pp. 1-26.

[29] G. A. Miller, "WordNet: a lexical database for English", *Communications of the ACM*, vol. 38, pp. 39-41, 1995.

[30] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer, "Extending VerbNet with novel verb classes," in *Proceedings of LREC*, 2006.

[31] J. Kim, Supervenience and mind: Selected philosophical essays: Cambridge University Press, 1993.

[32] R. W. Hamming, "Error Detecting and Error Correctiong Codes", Bell System Techincal Journal, vol. 29, no. 2, April 1950.

[33] I. Androutsopoulos and P. Malakasiotis, "A survey of paraphrasing and textual entailment methods," *Journal of Artificial Intelligence Research (JAIR)*, vol. 38, pp. 135-187, 2010.

News Events Prediction Based on Casual Inference in First-Order Logic (FOL)

Sina Dami[1*], Hossein Shirazi[2] and Ahmad Abdollahzadeh Barforoush[3]
1* - Corresponding Author: Department of Information Communication and Security, Malek-Ashtar University of Technology, Tehran, Iran.
2- Department of Information Communication and Security, Malek-Ashtar University of Technology, Tehran, Iran.
3- Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran.
[1*]Dami@mut.ac.ir, [2]Shirazi@mut.ac.ir, [3]Ahmad@ce.aut.ac.ir

Abstract- A novel method for future event prediction is proposed in textual environment. Proposed method is able to produce an event prediction model through generalization of cause events and then predict the effect events by using causal rules. First, the events of interest are extracted from domain-specific texts via an event representation model at semantic level, and are stored in the form of a graphical model in ontology as a posteriori (dynamic) knowledge. Then, a set of domain-specific causal rules in first-order logic (FOL) are fed into the machine as a priori (common-sense) knowledge. In addition to this common-sense knowledge, several large-scale ontologies containing DBpedia, VerbNet and WordNet are used for modeling contextual (static) knowledge and generalizing events. Finally, all types of these knowledge are integrated in a standard Web ontology Language (OWL) to perform causal inference. Empirical evaluation on real news articles showed that our method was better than the baselines.

Keywords- News Prediction, Event Representation, Semantic Processing, Ontology, Causal Inference, FOL.