استخراج جملات عمومی از ویکی پدیای فارسی:

تلاشی در جهت ساخت آنتولوژی دانش عرفی

مهدی مرادی[۱]، بهرام وزیرنژاد[۲]، پروانه خسروی زاده[۳] و هادی عبدی قویدل[۴]
[۱]دانشگاه صنعتی شریف، mehdi_moradi@mehr.sharif.ir
[۲] دانشگاه صنعتی شریف، bahram@sharif.ir
[۳]دانشگاه صنعتی شریف، khosravizadeh@sharif.ir
[۴]دانشگاه صنعتی شریف، hadi_abdighavidel@mehr.sharif.ir

چکیده - متون موجود در اینترنت حاوی انواع مختلفی از دانش هستند که با استفاده از تکنیک های مهندسی دانش می‌توان این نوع دانش ها را که در حوزه ی وب معنایی کاربردهای متنوع و مختلفی دارند استخراج نمود. دانش عرفی یکی از کاربردی‌ترین دانش‌ها در حوزه محاسباتی است که اغلب در قالب جملاتی عمومی مانند "پستانداران بچه زا هستند. " بیان می‌شود. ویژگی بارز این نوع جملات بر خلاف جملات غیر عمومی که اشاره به شخص، موجود یا شی خاص دارد بیان قاعده‌ای کلی راجع به انواع یا رخدادها است. هدف این تحقیق ارائه‌ی روشی خودکار جهت استخراج این نوع جملات از ویکی پدیای فارسی و استفاده از آن در ساخت آنتولوژی‌های دانش عرفی برای زبان فارسی است.

کلیدواژه‌ها : استخراج جملات عمومی، دانش عرفی، آنتولوژی، تشخیص الگو

۱. مقدمه

عصر ارتباطات و گسترش استفاده از نامه‌های الکترونیکی، پیام‌های فوری (IM)، اسناد، وبلاگ‌ها، مقالات خبری و صفحات خانگی، روزانه حجم گسترده‌ای از محتوی متنی را وارد وب می‌کند. اطلاعات متنی تمامی جوانب زندگی‌مان را در بر گرفته‌اند و نیاز به تکنولوژی‌های جدید برای مدیریت و پردازش این حجم عظیم اطلاعات روز به روز بیشتر احساس می‌شود. هر چند استفاده از روش‌های مبتنی بر کلمات کلیدی و روش‌های آماری در بازیابی اطلاعات، داده کاوی و سیستم‌های پردازش زبان طبیعی به موفقیت‌های دست یافته است ولی متخصصان این حوزه معتقدند که این روش‌ها تنها درکی سطحی از متون ارائه می‌دهند و برای پیشرفت در مدیریت اطلاعات، و دست‌یابی به درکی عمیق‌تر از متون، ماشین‌ها نیازمند دسترسی به حجم گسترده‌ای از دانش معنایی دنیای پیرامون خود هستند. وب معنایی که مهمترین هدفش را قادر ساختن ماشین به منظور تحلیل داده‌های موجود در وب (از محتوی و لینک‌ها گرفته تا تعاملات انسان- ماشین) می داند، هنوز نتوانسته است به اهدافش نایل شود [۱]. اساس و پایه‌ی وب‌معنایی آنتولوژی‌ها هستند. دانش معنایی مورد نیاز برای تفسیر منابع از این آنتولوژی‌ها بدست می‌آید. تاکنون بیشتر تلاش‌های صورت گرفته برای ساخت این آنتولوژی‌ها معطوف به توسعه آنتولوژی‌های حوزه‌های تخصصی مانند آنتولوژی ژن[1] بوده است که هدفشان بازنمایی دانش مربوط به همان حوزه تخصصی بوده است (برای مثال دانش مربوط به ژن و تولیدات ژنی). هرچند استفاده از URIها نقش مهمی در برقراری ارتباط مابین آنتولوژی‌های حوزه های مختلف داشته است، ولی در عمل ادغام کردن این آنتولوژی های کار دشواری است. مشکل اصلی عدم وجود یک آنتولوژی دانش عرفی است که بیشتر این آنتولوژی‌های حوزه‌ای متکی بر مفاهیم و روابط تعریف شده‌ی آن حوزه خاص از دانش بشری می‌باشند. برای مثال، واژه "دلستر" در آنتولوژی "مواد غذایی" تعریف شده است، ولی در آنتولوژی "رستوران" احتمالا رابطه "دلستر را در رستوران می نوشند" یا اینکه "تنها موادی را که مایع هستند میتوان نوشید و "دلستر مایع است" ذکر نشده است. از

[1] Gene Ontology (http://www.geneontology.org)

طرف دیگر، ساخت دستی این نوع آنتولوژی با توجه به حجم گسترده این دانش پرهزینه و زمانبر خواهد بود. در این تحقیق سعی شده با استفاده از تکنیک‌های زبان‌شناسی رایانشی این دانش (دانش عرفی که معمولا در قالب جملات عمومی بیان می شود) از ویکی پدیای فارسی استخراج گردد.

۲. دانش عرفی[۲] (حس عام):

از همان اولین سال‌های ظهور هوش مصنوعی، متخصصان این حوزه بر این مهم واقف بودند که هر سیستم هوشمندی نیازمند دسترسی به دانش عرفی است [۲]. برای ما انسان‌ها "دانش عرفی" تداعی کننده "قضاوت صحیح" است در حالی که "دانش عرفی" در حوزه زبان‌شناسی رایانشی[۳] اشاره به میلیون‌ها حقیقت کوچک و واضح دارد که برای انسان موضوعاتی پیش پا افتاده به حساب می‌آیند.
"سوزن تیز است."، "برای باز کردن در ابتدا باید دستگیره را چرخاند" و"اگر روز تولد کسی را فراموش کنید، از دستتان ناراحت خواهد شد". این جملات نمونه‌های این دانش هستند: دانش عرفی، گستره وسیعی از دانش دنیای پیرامون را شامل می‌شود و جوانبی از دانش فضایی، فیزیکی، اجتماعی را پوشش می‌دهد. به ما می‌گوید که چه غذایی خوردنی است و چه حیوانی خطرناک است[۳]. چون فرض بر این است که همه از این "دانش عرفی" بهره‌مندند، این نوع دانش معمولاً در ارتباطات اجتماعی محذوف است. برای داشتن درکی عمیق از اطلاعات متنی رایانه‌ها نیازمند حجم گسترده‌ای از این نوع دانش هستند که در حال حاضر فقط در اختیار انسان است. هدف ما استخراج و تزریق این نوع دانش به ماشین است.

۳. جملات عمومی[۴] :

بخش بزرگی از دانش عرفی ما در قالب جملات عمومی بیان می‌گردد. جملاتی مانند جملات زیر:
الف. فلزات جامدند.
ب. پستانداران بچه زا هستند.
به چنین جملاتی که قاعده‌ای عمومی برای مثال راجع به فلزات یا پستانداران بیان می‌کنند جملات عمومی گفته می‌شود. این نوع جملات همواره یکی از پیچیده‌ترین مسائل حوزه زبانشناسی صوری و یکی از موضوعات مطرح در زبانشناسی فلسفی بوده اند [۴]. جملات عمومی بایستی از جملات سور عمومی تمایز داده شوند برای مثال جمله "اسب‌ها چهار پا دارند" در شرایط عادی برای همه اسب‌ها صادق است ولی ممکن است اسب‌های یافت شوند که یک پایشان را از دست داده باشند ولی این اسب‌ها همچنان جزء طبقه اسب‌ها به حساب می‌آیند. در واقع جملات عمومی علی رغم وجود استثنائات، همواره صادق‌اند. همچنین بایستی بین این جملات و جملاتی که می‌شود از طریق استنتاج قاعده یا قانونی کلی از آنها استخراج کرد تمایز قائل شد. برای مثال، جمله "خورشید از شرق طلوع می‌کند." قاعده‌ای را بیان می‌کند در حالی که جمله "خورشید امروز ازشرق طلوع کرد." نمونه ای را بیان می‌کند که وقتی در کنار دیگر مثال‌های اینچنینی قرار می گیرد، می‌شود قاعده ای ازآن استخراج کرد [۵].
در مورد تفسیر صحیح این نوع جملات در زبان‌شناسی رایانشی و فلسفه زبان نظرات متفاوتی ارائه شده است. برای مثال: Cohen (۲۰۰۲) پیشنهاد می‌کند که جملات عمومی باید بصورت احتمالاتی تفسیر گردند[۶] در حالی که Carlson این نوع جملات را در ارتباط با رخدادهای عادی[۵] یا دانش طبقه‌ای[۶] می‌دانند [۴]. تفسیر صحیح جملات عمومی روش نمایش دانش[۷] را تحت تاثیر قرار می‌دهد.

۱,۳. بررسی جملات عمومی در زبان فارسی:

مبحث تفسیر جملات عمومی مباحث بسیاری را در زبان‌شناسی و فلسفه بدنبال داشته است. مسئله قابل تامل در زمینه چنین جملاتی، استثناء پذیر بودن آنها بدون تغییر ارزش صدقشان است. برای مثال جمله عمومی: سگ‌ها پارس می‌کنند بعلت عدم توانایی بعضی از سگ‌ها در پارس کردن غلط محسوب نمی‌شود و برعکس، با فرض اینکه تعداد معینی نمونه ویژگی مشترک P را دارند به معنی این نیست که جمله‌ای عمومی مبنی بر انتساب این ویژگی به آن نوع خاص وجود دارد: این مسئله وقتی این نمونه‌ها نادر باشند عیانتر خواهد بود.
از لحاظ زبان‌شناختی جملات عمومی حداقل در انگلیس و فارسی و بیشتر زبان‌ها (نه همه ی آنها) مشخصه صوری خاصی که آنها را ازجملات معمولی متمایز کند، ندارند [۷]. این نوع جملات از لحاظ روساخت نحوی، ساختواژی و آواشناختی هیچ تفاوتی با دیگر جملات زبان ندارند، با این حال این نوع جملات از لحاظ معنایی

[2] commonsense knowledge
[3] Computational linguistics
[4] Generic sentences
[5] habitual actions
[6] taxonomic knowledge
[7] Knowledge representation

دارای چند ویژگی خاص هستند: جملات عمومی (۱) جملاتی هستند ایستا (۲) که از لحاظ واژگانی گزاره‌ای غیر ایستا دارند و(۳) و جملاتی هستند تعریفی[۸] (در مقابل مصداقی[۹]) و غیر یکنوا[۱۰]. عدم وجود مشخصه‌ای خاص برای چنین جملاتی کار استخراج آنها را مشکل می‌کند. با این حال در بافت‌های خاصی برای مثال در پیکره های متنی مربوط به دانشنامه‌ها بعلت نوع دانشی که در این متون وجود دارد (دانشی کلی راجع به جهان پیرامون) با استفاده از الگوهایی خاصی که از توالی کلمات خاص مانند "همه‌ی" ،"هر"،"اغلب" و فعل‌هایی مانند "است، می شوند" تشکیل شده است می‌توان این نوع جملات را مشخص و استخراج کرد.

۴. مروری بر ادبیات:

با توجه به اهمیت دانش عرفی در وب معنایی، تحقیقات زیادی در زمینه باز نمایی این نوع دانش صورت گرفته است. بزرگترین پروژه در این زمینه ،پروژه Cyc است که تکمیل بانگ اظهاراتش[۱۱] دو دهه به طول کشیده است [۸]. پروژه OpenMind Commonsense (Singh ۲۰۰۲) دومین پروژه بزرگ در زمینه دانش عرفی است که برخلاف Cyc که بانک اطلاعاتیش توسط مهندسان دانش جمع‌آوری گشته، به کاربران معمولی اینترنت متکی بوده و این کاربران بودند که این نوع جملات را به پایگاه داده‌ی آن افزوده اند [۹]. این پروژه ۱/۶ میلیون عبارت عرفی جمع آوری کرده است که ۷۰۰۰۰۰ جمله قابل دسترسی از طریق پایگاه داده Conceptnet است. در conceptnet دانش عرفی در قالب جملات زبان طبیعی ذخیره می‌شود، ازاین رو بازنمایی روابط در قالبی صوری کار مشکلی خواهد بود. روش بهینه دیگر استفاده از ابزارهای پردازش زبان طبیعی جهت استخراج چنین حقایقی از منابعی همچون ویکی‌پدیاست که پایگاهی جامع از اطلاعات دنیای پیرامون ماست و بطور منظم توسط کاربرانش به روز رسانی و کنترل می‌شود.

۵. مواد و روش‌ها:

۵،۱. پیکره ی متنی:

ویکی‌پدیا به انگلیسی(Wikipedia) :یک دانشنامه اینترنتی چندزبانه با محتویات آزاد است که با همکاری افراد داوطلب نوشته می‌شود و مقالات آن می‌تواند توسط هر کسی که به اینترنت دسترسی دارد، ویرایش گردد. ویکی‌پدیای فارسی با داشتن حدود ۱۲۷ هزار مقاله در حوزه‌های مختلف پیکره متنی مناسبی جهت مطالعات زبانی می‌باشد. ویکی‌پدیا به کاربران این امکان را می دهد که نسخه کاملی از تمامی مقالات را در قالب یک فایل با فرمت Xml دانلود کنند. برای انجام این تحقیق نسخه 2011-04-10 این دانشنامه بارگذاری شده و توسط کتابخانه BeautifulSoup برچسب‌های XML جدا و پیکره به متن ساده[۱۲] تبدیل شد.

۵-۲. روش استخراج دانش:

تاکنون روشهای مختلفی برای استخراج دانش از متن پیشنهاد شده است. دو روش عمده برای این منظور، روش استخراج با دانش ضعیف (روشهای آماری) و روشهای غنی از دانش (استدلال منطقی) می باشند. در روش‌های آماری پردازش متن طریق محاسبه بسامد تکرار[۱۳]، هم وقوعی[۱۴] و باهم‌آیی[۱۵] کلمات صورت می گیرد درحالی که در روش دوم از روشهای نمادینی مانند روشهای منطقی مبتنی بر الگو و زبان-پایه استفاده می‌شود. روشهای زبان ـ پایه مانند تحلیل کامل نحوی، تحلیل ساختواژی ـ نحوی[۱۶] ، تجزیهٔ الگوهای لغوی ـ نحوی، پردازش معنایی و درک متن عموماً وابسته به زبان هستند و برای استخراج دانش از منابع غیرساخت یافته به کار می‌روند [۱۰]. یوچول جونگ و همکاران با استفاده از آنتولوژی دانش‌عرفی زبان انگلیسی و با کمک تکنیک های زبان‌شناسی رایانشی توانستند یک آنتولوژی برای زبان کره‌ای بسازند [۱۱]. روبرت اسپیر و همکاران در روشی جدید و خلاقانه با طراحی بازی هدف‌دار توانسته‌اند دانش عرفی را از کاربران این بازی جمع‌آوری کنند [۱۲]. در روشهای مبتنی بر الگو، با کمک الگوها و کلیدواژه‌های خاص روابط مورد نظر استخراج می‌شود .روش مورد استفاده در این تحقیق روش مبتنی بر الگو بوده است که برای اولین بار توسط هرست پیشنهاد شد [۱۳]. سعی شده است تا الگوهای مورد جستجو با توجه به روابط تعریف شده در ConceptNet انگلیسی (جدول ۱) استخراج شوند (جدول ۲). در

[8] intentional
[9] extensional
[10] non-monotonic
[11] Assertion

[12] Plain text
[13] Unigram
[14] Co-occurrence
[15] collocation
[16] morphosyntactic

جدول ۳ الگوهای تعریف شده برای زبان فارسی و نیز روابط معادلشان به همراه یک مثال نشان داده شده‌اند.

جدول ۱

روابط تعریف شده در ConceptNet

ردیف	رابطه
۱	DesireOf
۲	CapableOf
۳	MotivationOf
۴	PropertyOf
۵	LocationOf
۶	UsedFor
۷	IsA
۸	EffectOf
۹	CapableOfReceivingAction
۱۰	part of
۱۱	SubeventOf
۱۲	MadeOf
۱۳	DefinedAs
۱۴	FirstSubeventOf

جدول ۲

تعداد جملات استخراج شده در هر الگو

ردیف	نوع الگو	تعداد(جمله)
۱	خانواده	۴۳
۲	گفته می‌شود	۲۶۷۲
۳	استفاده می‌شود	۹۵
۴	همه	۴۱
۵	نوعی	۸۴۷
۶	ساخته می‌شوند	۱۹۴

جدول ۳

الگوهای استخراجی از ویکی پدیا فارسی و رابطه معادلشان در ConceptNet

ردیف / مثال	الگو	رابطه قابل تطبیق در ConceptNet
۱	X, Yی است از (خانواده ی|گونه ی|رده ی|راسته ی) z	IsA
مثال	لیکو پرنده‌ای است از خانوادهی لیکویان.	IsA
۲	X به Y گفته می‌شود.	IsA یا DefinedAs
مثال	آبگیر یا بِرکه به جایی گفته می‌شود که مقداری آب در آن جمع شده باشد.	
۳	از X برای Y استفاده می‌شود.	UsedFor
مثال	در خانه، استیک اسید رقیق برای باز کردن لوله ها استفاده می‌شود.	
۴	(همه ی|تمام|اکثر|بیشتر)Xها Y هستند.	IsA
مثال	همه پستانداران و پرندگان همچون انسان دارای دو بطن هستند.	
۵	X نوعی Y است که. ..	IsA
مثال	کچاپ نوعی سس است که معمولاً از گوجه فرنگی رسیده تهیه می‌شود.	
۶	Xها از Y ساخته می‌شوند.	MadeOf
مثال	تانکها از آلمینیوم یا فولاد ساخته می‌شوند.	

نتایج و پیشنهادها:

هرچند کارهای صورت گرفته جهت توسعه آنتولوژی‌های دانش عرفی فارسی برای وب معنایی بسیار اندک هستند ولی نتایج

جدول ۴

مقادیر مربوط به ارزیابی جملات استخراج شده

ردیف	الگو	تعداد کل جملات	تعداد جملات ارزیابی شدع	ارزیاب ۱	ارزیاب ۲	درصد ۱	درصد ۲	جمع کل
۱	خانواده	۴۳	۴	۴	۴	۱۰۰	۱۰۰	۱۰۰
۲	گفته می شود	۲۶۷۲	۲۶۷	۲۰۰	۱۹۰	۷۵	۷۱	۷۳
۳	استفاده می شود	۹۵	۹	۷	۹	۷۸	۱۰۰	۸۹
۴	همه	۴۱	۴	۳	۲	۷۵	۵۰	۶۳
۵	نوعی	۸۴۷	۸۴	۶۵	۷۰	۷۷	۸۳	۸۰
۶	ساخته می شوند.	۱۹۴	۱۹	۹	۸	۴۷	۴۲	۴۵
۷	جمع	۳۸۹۲	۳۸۷	۲۸۸	۲۸۳	۷۵	۷۴	۷۵

بدست آمده در این تحقیق نشان می دهد که استفاده از تکنیک های پردازش زبان طبیعی و تمرکز بر روی جملات عمومی نوید حوزه‌ای پویا را در فناوری‌اطلاعات می‌دهد. در این تحقیق ما با استفاده از این روش‌ها توانستیم از ۱۵۰ هزار مقاله فارسی ویکی‌پدیا ۳۸۹۲ جمله عمومی استخراج کنیم. بیشترین تعداد جملات مربوط به رابطه ۲ و کمترین مربوط به رابطه ۴ می‌باشد. برای ارزیابی میزان دقت این روش در استخراج این نوع جملات از دو گویشور زبان فارسی خواسته شد که نسبت به حس عام(بیان قاعده کلی) بودن یا نبودن ۱۰ درصد از جملات هر گروه که بصورت تصادفی انتخاب شده بودند قضاوت کنند(جدول ۴). نتایج بدست آمده نشان می دهد که روابط ۱ و ۳ بهترین و رابطه ۶ نامناسب ترین الگو برای استخراج جملات عمومی می باشد. برای بهبود عملکرد الگوهای استخراج در کارهای بعدی می توان از دیگر ابزارهای پردازش زبان طبیعی مانند برچسب زن اجزای کلام، چانکر (جهت مشخص کردن عبارات اسمی و فعلی) و نیز برای ارزیابی بهتر خروجی‌ها از پیکره‌های تگ خورده جملات عمومی استفاده کرد.

مراجع

[۱] T. Berners-Lee,and M. Fischetti, Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor. Britain: Orion Business. 1999.

[۲] J. McCarthy, PROGRAMS WITH COMMON SENSE, in Proceedings of the Teddington Conference on the Mechanization of Thought Processes, 756-91. London, 1959.

[۳] B. Smith. Formal ontology, common sense and cognitive science. International Journal of Human Computer Studies, 43(5):641–668, 1995.

[٤] G. N. Carlson and F. J. Pelletier, editors, The Generic Book, University of Chicago Press, 2005.

[٥] G. N Carlson, , Generic terms and generic sentences. Journal of Philosophical Logic, 11(2):145–181. 1982.

[٦] A . Cohen, Generics, frequency adverbs, and probability, Linguistics and Philosophy 22: 221-253. 1999.

[٨]D. Lenat, Cyc: Towards programs with common sense. Communications of the ACM, 8(33):30–49., 1990.

[٩]P. Singh, T. Lin; E. T. Mueller, G. Lim, T. Perkins and W. L Zhu, Open mind common sense: Knowledge acquisition from the general public. In CoopIS/DOA/ODBASE, 1223–1237, 2002.

[۷] کوروش صفوی، ,درآمدی بر معنی شناسی،ص ۱۴۹-۱۵۷، سوره مهر، ۱۳۸۷.

[۱۰] مهرنوش شمس‌فرد و احمد عبداله زاده بارفروش "استخراج دانش مفهومی از متن با استفاده از الگوهای زبانی و معنایی" مجله تازه‌های علوم شناختی، سال ۴، شماره ۱، ۱۳۸۱.

[۱۱] Y. Jung, J. Lee, Y. Kim, J. Park; Building a Large-Scale Commonsense Knowledge Base by Converting an Existing One in a Different Language Sung-Hyon Myaeng1, and Hae-Chang Rim, CICLing 2007, LNCS 4394, pp. 23 – 34, 2007.

[۱۲] R. Speer, C. Havasi and H. surana ;using verbosity: Common Sense Data from Games with a Purpose, Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010).

[۱۲] M. A Hearst, Automatic acquisition of hyponyms from large text corpora, Proceedings of the 14th conference on Computational linguistics-Volume 2, 539-545, 1992.