معرفی الگوهای واژگانی- نحوی
برای شناسایی رابطهٔ شمول معنایی در ویکی‌پدیای فارسی

دکتر پروانه خسروی‌زاده
دانشگاه صنعتی شریف
Khosravizadeh@sharif.ir

علی فارسی‌نژاد
دانشگاه صنعتی شریف
farsinejad@mehr.sharif.ir

چکیده

تشخیص الگو یکی از روش‌های استخراج دانش و کشف روابط میان مفاهیم زبانی است. بنابراین برای استخراج دانش مفهومی از میان داده‌های زبانی باید به طراحی و ساخت الگوهای معنایی پرداخت. مقاله حاضر ضمن بررسی روش‌های موجود مبتنی بر الگو به معرفی چند الگوی واژگانی- نحوی برای تشخیص رابطه شمول معنایی می‌پردازد. داده‌های لازم برای آزمایش الگوها از ویکی‌پدیای فارسی انتخاب شده است. این انتخاب به این دلیل صورت گرفته که ویکی‌پدیا به عنوان یک متن ساخت‌یافته، منبع خوبی برای استخراج روابط معنایی است. الگوهای معرفی شده در این نوشتار بر روی متون موجود در ویکی‌پدیا آزمایش شده و دقت هر الگو مورد ارزیابی قرار گرفته است.

کلیدواژه‌ها: شمول معنایی، روابط مفهومی، پردازش زبان طبیعی، الگوهای هیرست، تشخیص الگو.

معرفی الگوهای واژگانی– نحوی
برای شناسایی رابطه شمول معنایی در ویکی‌پدیای فارسی

پروانه خسروی‌زاده — **دانشگاه صنعتی شریف** — **Khosravizadeh@sharif.ir**

علی فارسی‌نژاد — **دانشگاه صنعتی شریف** — **farsinejad@mehr.sharif.ir**

چکیده

تشخیص الگو یکی از روش‌های استخراج دانش و کشف روابط میان مفاهیم زبانی است. بنابراین برای استخراج دانش مفهومی از میان داده‌های زبانی باید به طراحی و ساخت الگوهای معنایی پرداخت. مقاله حاضر ضمن بررسی روش‌های موجود مبتنی بر الگو به معرفی چند الگوی واژگانی- نحوی برای تشخیص رابطه شمول معنایی می‌پردازد. داده‌های لازم برای آزمایش الگوها از ویکی‌پدیای فارسی انتخاب شده است. این انتخاب به این دلیل صورت گرفته که ویکی‌پدیا به عنوان یک متن ساخت‌یافته، منبع خوبی برای استخراج روابط معنایی است. الگوهای معرفی شده در این نوشتار بر روی متون موجود در ویکی‌پدیا آزمایش شده و دقت هر الگو مورد ارزیابی قرار گرفته است.

کلیدواژه‌ها: شمول معنایی، روابط مفهومی، پردازش زبان طبیعی، الگوهای هیرست، تشخیص الگو

مقدمه

استخراج دانش از متن همواره در پردازش زبان طبیعی مطرح بوده است. در این میان، استخراج روابط مفهومی، اعم از روابط پایگانی و غیرپایگانی که برای توسعه هستان‌شناسی‌ها[1] به‌کار می‌روند، از اهمیت ویژه‌ای برخوردارند. استخراج روابط معنایی در بهینه‌سازی درخواست‌های بازیافت اطلاعات، پاسخ به پرسش در وب، ترجمه ماشینی و غیره نیز کاربرد دارد. این استخراج دانش از متن می‌تواند به شیوه‌های غیرنمادین (آماری) یا نمادین (منطقی یا زبان‌شناختی) و یا ترکیبی از هر دو رویکرد انجام شود. در روش تشخیص الگو که روشی زبان‌شناختی است، کلمات کلیدی از قبل تعریف شده و ساختارها یا الگوهایی که رابطه مورد نظر را نشان می‌دهند، مورد بررسی قرار می‌گیرند. در این مقاله دو روش برای استخراج رابطه شمول معنایی از ویکی‌پدیای فارسی دنبال می‌شود. الف) جستجوی الگوهای واژگانی- نحوی در متن مقالات ویکی‌پدیا.
ب) استخراج رابطه از ساختار صفحات ویکی.

[1] Ontology

1. شمول معنایی[۲]

رابطه شمول معنایی یکی از اساسی‌ترین روابط مفهومی در نظام زبان است. نتایج پژوهش شریفی و مولوی (1387) که به تشخیص پربسامدترین روابط مفهومی در داده‌های زبانی اختصاص دارد نشان می‌دهد بیشترین درصد پیوندهای واژگانی در زبان فارسی به روابط شمول معنایی، روابط نقشی، تقابل معنایی، باهم‌آیی و هم‌معنایی اختصاص دارد.

از سوی دیگر، به اعتقاد ارسطوپور و آزاد (1386) در نظام‌های اطلاع‌رسانی نیز که برچسب‌های موضوعی همان مقوله‌های مفهومی به‌شمار می‌روند، بیشترین میزان نارسایی‌های مقوله‌های مفهومی در حوزه سازمان‌دهی موضوعی مربوط به شمول معنایی است. لاینز[۳] (1993: ص 291) رابطه شمول معنایی را رابطه بین یک مفهوم خاص‌تر و جزئی‌تر، و یک مفهوم کلی‌تر و شامل‌تر می‌نامد. به عنوان مثال هر یک از واژه‌های گنجشک، کبوتر و عقاب نسبت به واژه پرنده واژه زیرشمول[۴] به‌حساب می‌آیند و نسبت به یکدیگر هم‌شمول[۵] محسوب می‌شوند. پرنده نیز واژه شامل[۶] آنهاست. صفوی نیز شمول معنایی را رابطه میان دو واژه‌ای می‌نامد که "یکی از چنان وسعتی برخوردار است که معنی واژه دیگر را نیز شامل می‌شود. در این سلسله مراتب، واژه‌ها یا فراگیرنده و یا هم‌شمول خواهند شد" (صفوی، 1380).

تصور ما از شمول دو گونه می‌تواند باشد. یا تصورمان از شمول این است که مصداق واژه زیرشمول، زیرمجموعه مصادیق واژه شامل است، یا قضاوتمان این است که مولفه‌های معنایی واژه زیرشمول، زیرمجموعه مولفه‌های معنایی واژه شامل است. به عنوان مثال، اگر X مجموعه گل‌ها و Y مجموعه شقایق‌ها باشد، آنگاه Yزیرمجموعه محض X است. تعریف مولفه‌های معنایی نیز بدین ترتیب است که X زیرشمول Y است اگر و فقط اگر مولفه‌های معنایی که X را تعریف می‌کنند، زیرمجموعه محض مولفه های معناییY باشند. یکی از مشکلات این دو تعریف زیرمجموعه‌ای از رابطه شمول معنایی این است که چنین تعریفی روابط مفهومی بسیاری را مجاز می‌شمارد که گویشور زبان آنها را جزو تعاریف طبیعی رابطه شمول معنایی نمی‌داند. به‌عنوان مثال، اگرچه اسب نوعی حیوان است و در شمول معنایی واژه حیوان قرار می‌گیرد، اما ملکه که طبق این تعریف زیرشمول زن محسوب می‌شود، آیا واقعا *نوعی* زن است؟ کروز[۷] چنین روابطی را روابط رده‌ای[۸] می‌نامد و آنها را نوع خاصی از شمول معنایی می‌خواند (کروز، 1997، ص 137).

لاینز (1993: صص 291-302) استلزام یک‌طرفه را راهی برای آزمایش رابطه شمول معنایی می‌داند. بدین‌ترتیب که صحت یک جمله با واژه زیرشمول مستلزم صحت همان جمله با واژه شامل است و نه برعکس. برای نمونه، صحت جمله "این یک لاله است" مستلزم صحت جمله "این یک گل است" خواهد

---

[2] hyponymy
[3] Lyons
[4] hyponym
[5] co-hyponym
[6] superordinate

[7] Cruse
[8] taxonomy

بود ولی عکس آن صادق نیست (صفوی، 1379، ص 102). مزیت چنین تعریفی آن است که دانستن مولفه‌های معنایی واژه‌های درگیر در این رابطه ضروری نیست. هر چند این تعریف نیز کاستی‌های خود را دارد و همیشه رابطه شمول معنایی و استلزام درکنار یکدیگر قرار نمی‌گیرند. به‌طور مثال جمله زنبور "قوزکم" را نیش زد، مستلزم آن است که زنبور "پایم" را نیش زده باشد، درحالیکه "قوزک" زیرشمول "پا" نیست.

مشکل تعاریف یادشده این است که بیشتر توصیفی هستند تا تبیینی و دقیقا مشخص نمی‌کنند در معنی دو کلمه چه نهفته است که رابطه شمول معنایی را میان آن دو برقرار می‌کند.

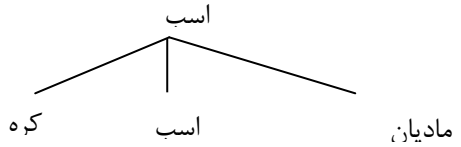
2. ویژگی‌های رابطه شمول معنایی

- رابطه شمول معنایی رابطه‌ای بازتابی[9]، گذرا[10]، اما نامتقارن[11] است. ویژگی بازتابی یعنی هر واژه شامل معنایی خودش به حساب می‌آید. گذرا یعنی $if\ xHy\ and\ yHz \Rightarrow xHz$ (که xHy به معنی x زیرشمول y است). به‌طور مثال زنبور نوعی حشره است و حشرات از جانداران هستند، بنابراین زنبور زیرشمول جانداران است. اما این رابطه متقارن و دوطرفه نیست. در واقع شکل دوطرفه این رابطه همان رابطه هم‌معنایی است.

- رابطه شمول معنایی در همه سطوح از واژه تا جمله جریان دارد. اما در سنت مطالعات روابط معنایی، رابطه شمول معنایی بیشتر میان عناصر واژگانی مانند اسم، فعل و صفت بررسی می‌شود. این رابطه از نوع مقوله‌ای است و دو عنصر واژگانی درگیر این رابطه باید متعلق به یک طبقه دستوری باشند.

- هرچند معمولا در واژگان هر زبان رابطه شمول معنایی به‌وفور دیده می‌شود، اما همواره برای هر واژه، یک واژه شامل وجود ندارد. لاینز (1993:ص 300) به واژه‌ای در زبان یونانی باستان اشاره می‌کند که شامل مشاغل بسیاری ازجمله نجاری، پزشکی، نوازندگی و کفاشی بوده است، در حالیکه امروزه چنین معادلی وجود ندارد.

- یک واژه چندمعنایی، بسته به اینکه در کدام مفهومش به کار رفته باشد، می‌تواند بر خودش شامل باشد و در چند جای سلسله مراتب شمول معنایی حضور یابد. واژه "اسب" در زبان فارسی واژه شاملی است که مفهوم سه واژه "مادیان"، "کره" و" اسب" را در بر می‌گیرد (صفوی، 1379، ص 101).



- این رابطه در هر زبانی با توجه به نظام طبقه‌بندی و نگاه سخنگویان آن زبان به جهان خارج تغییر می‌کند.

[9] reflective
[10] transitive
[11] asymmetric

3. روش‌های مبتنی بر الگو

در روش‌های مبتنی بر الگو ، واژه‌ای کلیدی و یا الگویی خاص انتخاب می‌شود و جستجو در اطلاعات درونداد، که معمولا به صورت فایل متنی است، برحسب آن واژه کلیدی یا الگوی خاص می‌گیرد. این الگوها انواع مختلفی اعم از نحوی یا معنایی، و عمومی یاخاص دارند و بر اساس نمونه، نحوه ساخت و نوع دانش مورد استخراج با هم متفاوتند. اولین بار لاینز (1993: ص 301) اشاره کرد که فقط مفاهیم شمول و زیرشمول می‌توانند در عبارتی مانند "پیتزا و دیگر خوراکی‌ها" بیایند و مثلا نمی‌توان گفت: "پیتزا و دیگر نوشیدنی‌ها". این الگوها از آن جهت برای لاینز مهم بودند که معتقد بود ما در مرحله یادگیری زبان، شمول‌ها و زیرشمول‌ها را از این طریق می‌آموزیم. در واقع بدون دانستن معنی و مولفه‌های معنایی کلمه "بانیان"، از عبارت "بانیان و درختان دیگر" می‌آموزیم که "بانیان" نوعی درخت است.

در قلمروی استخراج اطلاعات نیز روش‌های تطبیق الگو کاربرد وسیعی دارد. آلزوود و ایونز[12] (1988) با ارایه چند الگو و آزمودن آن‌ها بر روی داده‌های فرهنگ "وبستر" تعداد قابل ملاحظه‌ای از نمونه‌های رابطه شمول معنایی را استخراج کرده‌اند. اما اولین کسی که به بررسی عملی الگوها در متن نامحدود دست زد، مارتی هیرست[13] بود که در مقاله "اکتساب خودکار شمول معنایی از پیکره بزرگ متنی" (هیرست،1992) نشان داد با درک ساده‌ای از متن می‌توان به اطلاعات بسیاری دست یافت. به عنوان مثال با استفاده از الگویی مانند NP_0 such as $\{NP_1, NP_2 \ldots (and|or)\}$ NP_n

نتیجه زیر به دست آمده است:

For all $NP_i, 1 \leq i \leq n$ Hyponym(NP_i, NP_0)

هیرست بر این اساس و با بهره‌گیری از سه شرط زیر، یک مجموعه الگو برای استخراج رابطه شمول معنایی معرفی کرد. این الگوها باید از سه شرط زیر پیروی کنند:

1. با تکرار زیاد و در گونه‌های مختلف متن رخ دهند.
2. تقریبا همیشه رابطه مورد نظر را نشان دهند.
3. بدون نیاز به دانش رمزگذاری شده پیشین قابل تشخیص باشند.

الگوهای پیشنهادی هیرست به شرح زیر است:

1. *NP_0 ... such as*{*NP_1, NP_2 …* (*and*|*or*)}*NP*,,
2. *such NP as* {*NP*,} * {or|*and*} *NP*
3. *NP*{, *NP*} * {,} *or other NP*
4. *NP*{, *NP*} * {,} *and other NP*
5. *NP*{,} *including* {*NP* * {or|*and*} *NP*
6. *NP*{,} *especially* {*NP*,} * {or|*and*} *NP*

[12] Ahlswede & Evens
[13] Marti Hearst

ترجمه این الگوها در زبان فارسی اینگونه‌اند. در الگوهای 2،1،5 و 6 گروه اسمی X شامل معنایی گروه‌های اسمی شماره‌دار و گروه اسمی Y است، در 3 و 4 برعکس. (علامت ستاره به معنی تکرار است)

1. گروه اسمی X **مانند** {گروه اسمی 1، گروه اسمی 2 ...(و/یا)} گروه اسمی Y
2. **چنین** گروه اسمی X **مانند** **{**گروه اسمی 1،} * {و|یا} گروه اسمی Y
3. گروه اسمی X **یا دیگر** گروه اسمی Y
4. گروه اسمی X **و دیگر** گروه اسمی Y
5. گروه اسمی X **شامل** **{**گروه اسمی 1،} * {و|یا} گروه اسمی Y
6. گروه اسمی X **به‌ویژه** **{**گروه اسمی 1،} * {و|یا} گروه اسمی Y

شمس‌فرد (1381: صص 48-66) الگوهای هیرست را در فارسی آزموده است. الگوهای وی برای یافتن رابطه شمول معنایی، اقتباس لفظی و معنایی از الگوهای هیرست هستند. او برای ساخت نیمه‌خودکار هستان‌شناسی فارس‌نت (2010) نیز از این الگوها استفاده می‌کند.

الگوهای زبان فارسی

الگوهای هیرست درصورتیکه صرفا به همین شکل ترجمه شوند و در داده‌های فارسی به‌کار روند، دقت وفراوانی وقوع بالایی در زبان فارسی نشان نمی‌دهند. در این میان الگوهای 1 و 4 قابل تعمیم به داده‌های فارسی است، اما الگوهای دیگر از کارایی ضعیفی برخوردارند. از آنجا که هر الگو زیرمجموعه‌ای از روابط را نشان می‌دهد، این روابط باید تا حد امکان با جزییات بیشتری تعریف گردد. از این رو در این نوشتار، الگوهای زیر به عنوان نسخه تکمیلی الگوهای شمس‌فرد (2010) ارایه می‌گردد. برخی از الگوهای زیر با الگوهای هیرست و شمس‌فرد اشتراکاتی دارند و برخی نیز جدید هستند.

الگوی 1 بسط الگوی اول از مجموعه الگوهای هیرست است. الگوی 2 معادل الگوی هیرست است که دقتش آزمایش شده است. الگوی 3، 4، 5 و 6 جدید هستند. الگوی "عبارتند از" الگویی با تکرار زیاد و دقت نه‌چندان زیاد است. در ضمن ضمایر برگردانده شده توسط این الگو احتیاج به رفع ابهام دارند که معمولا مرجعشان عنوان مقاله ویکی است. به طور مثال در مقاله "کانادا"، عبارت یافته شده "شهرهای این کشور عبارتند از: مونترال ... " ضمیر این به عنوان اصلی مقاله، یعنی واژه "کانادا" برمی‌گردد.

	الگو	مثال
1	NP {هایی|اتی|انی} {مانند|مثل|همانند|چون|همچون|نظیر|از قبیل} NP_1 ،NP_2 ،... (و|یا) NP_n	شهرهایی مثل دالاس زنانی مانند الیزابت
2	NP و|یا دیگر NPها	آبگوشت، برنج و دیگر غذاها

3	این NP {ها|ات|ان}عبارتند از: NP_1، NP_2، ... ... NP_n	این روستاها عبارتند از: دهک
4	NP_1 (یکی از| از) NP ها|ان|ات است	ابن سینا یکی از دانشمندان..
5	NP NP ای/ی است	بامیه گیاهی است
6	NP_1 (نوعی|نام نوعی|گونه‌ای) NP است	سیتار نام نوعی ساز است

الگوهای اولیه با شم زبانی و درون‌نگری پیدا می‌شوند. هیرست روالی برای یافتن الگوهای بیشتر پیشنهاد می‌کند:

1. یک رابطه شمول معنایی را در نظر بگیرید، مثلا کشور و ایران.
2. فهرستی از واژه‌هایی پیدا کنید که چنین رابطه‌ای با هم دارند. این فهرست را می توان با کمک الگوهای دست‌ساز پیدا کرد. مانند کشور و فرانسه.
3. جاهایی را در پیکره بیابید که این عبارات در یک بافت مشترک آمده باشند و آن بافت را ذخیره کنید.
4. شباهت‌های بین این بافت‌ها را پیدا کنید و از مشترکات آنها که رابطه را نشان می‌دهند الگوی جدیدی بسازید.
5. وقتی یک الگوی جدید پیدا شد، از آن برای جمع‌آوری نمونه‌های جدید از رابطه استفاده کنید و به مرحله 2 بروید.

4. آزمایش:

ساختار یک زبان به شیوه‌های مختلف می‌تواند یک رابطه معنایی را نشان دهد، اما باید ساختارها و الگوهایی را یافت که با تکرار و دقت زیاد رابطه مورد نظر را نشان دهند. با این روش بسیار ساده نتایجی بااهمیت به‌دست می‌آید که می‌توان با استفاده از روش‌های مکمل آنها را بهبود بخشید.

ویکی‌پدیای فارسی به عنوان پیکره

ویکی‌پدیا دانشنامه‌ای همگانی و آزاد است که کاربران آن از سراسر جهان، می‌توانند به نوشتن و ویرایش نوشتارهای موجود در آن بپردازند و به رشد آن کمک کنند. ویکی‌پدیا با بیش از 150 هزار مقاله به زبان فارسی، هم به عنوان دایره‌المعارفی ساخت‌یافته در خط تعریفش و هم به عنوان متنی نامحدود در دیگر بخش‌ها، منبع مناسبی برای پردازش زبان فارسی به حساب می‌آید. نسخه قابل بارگذاری و ایستای صفحات، هم به صورت یک فایل بزرگ XML و هم به صورت فایل‌های فشرده به قالب b2z در دسترس است.

برنامه پس از جداکردن متن فارسی ویکی‌پدیا از برچسب‌ها، به یافتن الگوهای 1الی 6 که به صورت عبارات منظم یا باقاعده (regular Expression) دستی نوشته شده‌اند می‌پردازد. این الگوها چنانکه از نامشان

مشخص است هم بخش واژگانی و هم نحوی را در بر می گیرند. در این آزمایش الگوهای واژگانی بررسی شده است.

معمولا اولین سطر هر مقاله ویکی‌پدیا (سطر تعریف) بهترین سطر برای یافتن یک رابطه شمول معنایی است. با بررسی سطر اول مقالات ویکی مشخص شد که تقریبا همه آنها یکی از این الگوهای واژگانی- نحوی را در خود دارند. بعد از مراجعه به "راهنمای سبک" ویکی معلوم شد که این راهنما با ذکر اینکه "عنوان مقاله باید در اولین سطر بیاید" کاربر را ملزم می‌کند که ناخواسته از یک طرح رده‌بندی در ذهن خود و اولین بخش نوشته‌اش پیروی کند.

لازم به ذکر است که برای جستجوی دقیق‌تر این الگوها درجمله، نیاز به ابزارهای پردازش پایه زبان مانند برچسب‌زن اجزا کلام و تجزیه‌گر نحوی است که مرز گروه‌های اسمی را مشخص می‌کنند. از آنجا که چنین ابزاری برای زبان فارسی در دسترس نیست، نتایج باید با پَس‌پردازش دستی اصلاح شوند. حذف کلمات است، گروه‌های اضافه‌ای و جایگزینی گروه‌های اسمی بزرگ با هسته‌شان از این دست هستند.

تشخیص روابط با استفاده از ساختار صفحات

در ویکی‌پدیا ساختار صفحات به اندازه متن مقالات مهم است. صفحات ویکی که ساخت‌یافته‌تر از یک پیکره معمولی هستند، با برچسب‌های XML، امکان استخراج روابط معنایی مختلف، از جمله رابطه شمول معنایی را فراهم می‌کنند. به‌عنوان مثال، در یک فهرست گلوله‌ای عنوان فهرست شامل معنایی اقلام لیست است.

رده‌بندی: تمام مقالات ویکی‌پدیا توسط کاربران ویکی و با منطق سلیم بشری در قالب رده‌های مفهومی دسته‌بندی شده‌اند. این پیوندهای رده‌بندی در انتهای هر صفحه که طبقه‌بندی مفهومی عنوان مقاله را نشان می‌دهند، بخش مرتبط به کار ما هستند. فرض آزمایش‌نشده این است که عنوان هر مقاله و رده‌بندی‌هایش منبع خوبی برای استخراج رابطه شمول معنایی است. مثلا در مقاله‌ای با عنوان "دکارت" می‌توان رابطه شمول معنایی زیر را دید.

رده‌های صفحه: رنه دکارت | ریاضی‌دانان اهل فرانسه | فیلسوفان اهل فرانسه | زادگان ۱۵۹۶ (میلادی) | درگذشتگان ۱۶۵۰ (میلادی)

اما صفحه‌های رده‌بندی فقط محدود به رابطه شمول معنایی نمی‌شوند: به‌عنوان‌مثال در رده‌بندی مقاله "فوتبال"، تیم فوتبال ذکر شده است. اغلب این رده بندی‌ها روابط دیگری را مانند جزءواژگی در بر می‌گیرند که باید آنها را فیلتر کرد. ویکی‌پدیا برای بررسی رده‌های هر مقاله، فایل رده‌بندی مقالات را به‌صورت جداگانه ضمیمه کرده است. بنابراین، در این تحقیق، نیازی به تجزیه و جداکردن پیوندهای رده‌ها از متن نبود.

کاستی‌های احتمالی:

"افت اطلاعاتی" اصطلاحی است که در حوزه بازیابی اطلاعات و برچسب‌گذاری موضوعی مطرح می‌گردد. این اصطلاح بیانگر کاستی‌های پردازش است که می‌تواند به دلایل مختلف ازجمله خطای برنامه ویا استانده نبودن داده‌های مورد بررسی باشد. دراین تحقیق تلاش شده تا دقت لازم درالگوهای پیشنهادی اعمال گردد.

انتخاب منبع آزمایش الگوها نیز با تامل صورت گرفته است. مسلما نتایج بدست آمده باید در محک آزمون و پردازش دستی نیز قرار گیرند تا خطاهای احتمالی برنامه خود را نشان دهند.

تحلیل نتایج:

	الگو	تعداد تکرار			درصد دقت
		الگو	تکرار	دقت به درصد	
1	NP {هایی|اتی|انی} {مانند|مثل|همانند|چون|همچون|نظیر|از قبیل} NP_1، NP_2 ،.... (و|یا) NP_n	اتی مانند	257	68.9	75
		هایی مانند	444	91	
		انی مانند	480	57	
		اتی چون	210	65	
		هایی چون	329	94	
		انی چون	823	88	
		اتی مثل	211	63	
		هایی مثل	173	73	
		انی مثل	304	72	
2	... و|یا دیگر	143			70
3	این NP NP {ها|ات|ان}عبارتند از: NP_1، NP_2 ، NP_n	1346			69
4	NP_1 (یکی از| از) NP ها|ان|ات ست	820			80
5	NP NPاست	973			90

6	NP_1 (نوعی|نام نوعی|گونه ای) NP است	80	94

ارزیابی:

رویه ارزیابی، شم زبانی گویشور در انتخاب روابط مناسب شمول معنایی از بین نتایج بوده است. ادعای گزافی است که با چند الگوی ساده می‌توان این رابطه را مدل کرد، اما کیفیت نتایج تقریبا بالاست. به دلیل نوع متن، در بین زیرشمول‌های یافت شده، اسامی خاص به‌ویژه اسامی کشورها، شهرها و اشخاص فراوان بود. در بعضی از موارد، واژه‌های شامل در بالای سلسله مراتب قرار داشتند و شامل‌ها بلافصل واژه نبودند، مانند "موضوعات" و "کارها". در برخی موارد نتایج دربرگیرنده رابطه جزءواژگی نیز می‌شدند و در برخی موارد رابطه مبهم بود.

دلیل دقت بالای این الگوها می‌تواند اعمالشان روی متن ساخت‌یافته ویکی باشد. بعضی از الگوها تکرار زیاد و دقت کم و بعضی تکرار کم و دقت بالا دارند. به‌نظرمی‌رسد هر چقدر الگو را محدودتر کنیم روابط کمتر، اما دقیق‌تری بدست خواهد آمد و هر قدر الگو را عمومی‌تر کنیم روابط بیشتر اما با دقت کمتری خواهیم داشت.

نتیجه‌گیری:

سابقه پردازش‌های معنایی در زبان فارسی بسیار محدود است. ویکی‌پدیا به عنوان یک هستان‌شناسی ساخت‌یافته می‌تواند پیکره مناسبی برای آزمایش هم الگوریتم‌های آماری، و هم قواعد و الگوهای زبانی باشد. در نوشته حاضر، سعی بر آن بوده تا با استفاده از متن مقالات ویکی‌پدیای فارسی و ساختار صفحات آن، راهی ساده برای استخراج نیمه‌خودکار رابطه شمول معنایی ارائه شود. آزمون تجربیِ الگوهای پیشنهادی دقت قابل قبولی را نشان داده‌است. الگوهای پیشنهادی در این نوشتار به عنوان نسخه تکمیلی الگوهای شمس‌فرد (2010) ارائه شده است. برخی از این الگوها با الگوهای هیرست و شمس‌فرد اشتراکاتی دارند و برخی نیز کاملا جدید هستند.

منابع:

ارسطوپور، شعله و آزاد، اسدالله (1386)، نظریه برچسب‌گذاری و برچسب‌های موضوعی در سازمان‌دهی اطلاعات: نگاهی تطبیقی از زاویه ارتباطات متقاعدگرایانه، *فصلنامه کتابداری و اطلاع‌رسانی*، شماره چهارم، جلد دهم، صص 86-65.

ژورافسکی، دانیل و جیمز مارتین (1384)، معناشناسی واژگانی در معنا شناسی و بازیابی اطلاعات: هفت گفتار، ترجمه جعفر مهراد و محمدرضا فلاحت فومنی، شیراز: کتابخانه رایانه‌ای یا کتابخانه منطقه علوم و تکنولوژی. صص 102 – 41.

شریفی، شهلا و مولوی وردنجانی، آرزو (1387)، پربسامدترین روابط مفهومی میان واژگان، *پژوهشنامه ادب غنایی، مجله زبان و ادبیات دانشگاه سیستان و بلوچستان*، سال ششم، شماره دهم.

شمس‌فرد، مهرنوش و عبدالله‌زاده بارفروش، احمد (1381)، استخراج دانش مفهومی از متن با استفاده از الگوهای زبانی و معنایی، *تازه‌های علوم شناختی*، سال 4، شمارهٔ 1،صص 66-48.

صفوی، کورش (1379)، درآمدی بر معنی شناسی، پژوهشگاه فرهنگ و هنر اسلامی.

صفوی، کورش (1380)، نگاهی به مسئله شمول معنایی، گفتارهایی در زبانشناسی، تهران، انتشارات هرمس، صص 198-191.

Ahlswede, T. & Evens, M. (1988), Parsing vs. Text Processing in the Analysis of Dictionary Definitions, *ACL 88 Proceedings of the 26th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA.

Cruse, D.A. (1997), *Lexical semantics*, Cambridge, Cambridge University Press.

Hearst, M. (1992), Automatic Acquisition of Hyponyms from Large Text Corpora, *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France.

Lyons, J. (1993), *Semantics*, vol. 1, Cambridge, Cambridge University Press.

Shamsfard, M. (2010), Lexico-syntactic and Semantic Patterns for Extracting Knowledge from Persian Texts, *International Journal on Computer Science and Engineering*, Vol. 02, No. 06, pp. 2190-2196.

Introducing a few lexico-syntactic patterns to identify hyponymy relationship in Persian Wikipedia

Pattern recognition is one of the methods of extracting knowledge and exploring the relationship between semantic concepts. Therefore, to extract knowledge from the data, one must design and introduce semantic patterns.
This article reviews the existing methods based on multi-pattern model and introduces a few lexico-syntactic patterns for identifying hyponymy relationship. The data required for testing the patterns is selected from the Persian Wikipedia. Wikipedia as a structured ontology can be suitable for testing both, the statistical algorithms, and language rules and patterns. In this regards, patterns introduced in this article is carefully tested and evaluated on the documents of Persian Wikipedia. The research procedure of this paper is based on two methods for extracting hyponymy relationship among words; a) Searching lexico-syntactic patterns among Wikipedia's articles, and b) extracting the hyponymy relationship from the structure of wiki pages.
Patterns presented in this paper can be regard as a complementary version for Shamsfard's patterns (2010). A couple of them are somehow similar to those Hearst (1992), Shamsfard (2010) suggested, and the others are new.