«سایه‌نما»: سامانه درک استعاره با استفاده از ویژگی‌های معنایی مشترک بین دو واژه موجود در بافت استعاری

هادی عبدی قویدل[۱]
افشین رحیمی[۲]
پروانه خسروی‌زاده[۳]

چکیده

مقاله حاضر به معرفی سامانه‌ای با عنوان سایه‌نما می‌پردازد. ایدهٔ کلی این سامانه که در زمینه استعاره برای زبان فارسی است برای نخستین بار طرح و پیاده‌سازی می‌شود. سایه‌نما با هدف کمک به درک استعاره به صورت خودکار، عملیات یافتن ویژگی‌های معنایی مشترک بین دو واژه‌ای که در بافت استعاری حضور دارند انجام می‌دهد. فرایند کلی سایه‌نما بدین ترتیب است که نخست ویژگی-های معنایی مشترک بر اساس اطلاعات نقطه‌ای مشترک در هم‌رخدادی مرتبه دوم استخراج و در پی آن صفاتی معرفی می‌شوند تا بتوانند این نوع ویژگی‌ها را توصیف کنند. کاربرد این روش در سامانه‌های درک متن و به خصوص درک استعاره به صورت خودکار، کارایی سامانه را به طور قابل توجهی بهبود می‌دهد.

واژه‌های کلیدی: سامانه سایه‌نما، ویژگی‌های معنایی، اطلاعات نقطه‌ای، هم‌رخدادی مرتبه دوم

۱- مقدمه

مکالمات روزمره ما آکنده از مفاهیم پیچیده‌ای به نام استعاره است که بشر به طور خلاقانه در ارتباطات خود از آن استفاده می‌کند. این مفاهیم پیچیده در فرهنگ و زبان‌های خاص ابداع می‌شوند و در گنجینه مفاهیم پنهان آن زبان قرار می‌گیرند. دلیل پنهان بودن استعاره‌ها در زبان، بدین خاطر

[۱]- دانشجوی کارشناسی ارشد زبان‌شناسی رایانشی، مرکز زبان‌ها و زبان‌شناسی، دانشگاه صنعتی شریف
hadi_stlt@yahoo.com
[۲] - دانشجوی کارشناسی ارشد زبان‌شناسی رایانشی، مرکز زبان‌ها و زبان‌شناسی، دانشگاه صنعتی شریف
afshinrahimi@gmail.com
[۳]- استادیار، گروه زبان‌شناسی رایانشی، مرکز زبان‌ها و زبان‌شناسی، دانشگاه صنعتی شریف
khosravizadeh@sharif.edu

است که نشانه‌های زبانی خواه به صورت واژه و خواه به صورت جمله در مفهومی غیر صریح نمود می‌یابند.

در نتیجه تولید زبانی از دید استعاره، درکی نیز به مثابهٔ آن شکل می‌گیرد. انسان در مرحله‌ای این پدیده را خلق و در مرحله‌ای دیگر انتظار درک موفق مخاطب خود را دارد. اما با پا نهادن به دنیای فن‌آوری و خودکار کردن سخن، لازم است تا جامعه زبانی گسترش یابد. از این رو رایانه نیز در پاسخ به محرک‌های استعاری و تفسیر آن نقش مهمی در این جامعه دارد. برای به وقوع پیوستن این امر، رایانه باید به وسیله الگوریتم‌های زبان‌شناسی رایانشی و پردازش زبان طبیعی در زمینه استعاره آموزش ببیند.

برای زبان انگلیسی و برخی زبان‌های دیگر، الگوریتم‌ها به شیوه‌های مختلف به کار بسته شدند، اما به این زمینهٔ تحقیقاتی در ایران کمتر توجه شده و مطالعات کمی در این رابطه صورت پذیرفته است. نخستین تحقیق در این باره توسط عبدی و همکاران(۱۳۹۱) در دومین هم‌اندیشی زبان‌شناسی رایانشی معرفی شد. در این تحقیق، تفحص‌های رایانشی آغاز و روش‌های تشخیص خودکار و رفع ابهام معنایی مفاهیم استعاری و دارای مجاز بررسی شد. نتیجه این شد که اگر قابلیت پردازش مفاهیم استعاری و دارای مجاز به روش‌هایی مانند بازیابی اطلاعات، ترجمه ماشینی، استخراج اطلاعات، پاسخگوی خودکار، دسته‌بندی متون، درک گفتار، خوشه‌بندی متون، خلاصه‌سازی و تفسیر متون و همچنین عقیده‌کاوی یا تحلیل مقصود در زبان فارسی اضافه شود، دقت عملکرد و میزان کارایی آنها به مقدار قابل توجهی افزایش خواهد یافت.

در گام دوم این تحقیق، زبان فارسی در بوته آزمایش قرار گرفته است. به نقل از بایلیس[1] (۲۰۱۱)موسسه آمریکایی موسوم به یارپا[2] در حال انجام پروژه‌ای است که قرار است طی آن برای تجزیه و تحلیل مسائل فرهنگی و اتخاذ سیاست‌های بهتر، نرم‌افزاری برای تجزیه و تحلیل زبان استعاری تهیه شود.

در تحقیق حاضر، فرایند کلی، یافتن راه حلی برای استخراج ویژگی‌های معنایی مشترک بین دو واژه است. از آنجا که پدیده استعاره معمولاً زمانی ساخته می‌شود که مفهومی از زاویه دید ویژگی‌های مفهومی دیگر نگریسته شود و این زاویه دید به منزله ویژگی‌های مشترک و مجاورت معنایی دو واژه است، یافتن این ویژگی‌های مشترک گام موثری در توانایی درک خودکار استعاره است. این پژوهش بر اساس روش اطلاعات نقطه‌ای مشترک در هم‌رخدادی مرتبه دوم بر روی پیکره همشهری انجام گرفته و منجر به تشکیل سامانه درک استعاره به نام «سایه‌نما» شده است. به عنوان مثال سامانه «سایه‌نما» موفق شده است در پیکره همشهری، ویژگی برجستهٔ میان واژه‌هایی نظیر انسان و روباه را مکار و حیله‌گر بیابد.

---

[1]-Bayliss,

[2]-US Intelligence Advanced Research Project Activity

۲- پیشینه مطالعات

به طور کلی درک استعاره به دو روش مختلف صورت می‌گیرد. در روش اول مفهوم مبدا به عنوان نماینده یک طبقه از مفاهیم مطرح می‌شود که دارای مشخصه معنایی خاص مشترکی با مفهوم مقصد است. به عنوان مثال در جمله «علی گرگ است» گرگ به عنوان نماینده طبقه‌ای از موجودات مطرح می‌شود که دارای مشخصه معنایی «بی‌رحمی» است. بنابر این درک جمله به صورت مستقیم و با استفاده از معنای ثبت شده برای گرگ به عنوان «موجودی بی‌رحم» انجام می‌پذیرد. در پژوهش‌های جنتنر[1](۱۹۸۳) مطرح شده است که این روش برای درک مفاهیم استعاری جدید به کار نمی‌رود بلکه در صورتی که مفهوم استعاری به دفعات مکرر به کار برده شود مفهوم مبدا می‌تواند به عنوان نماینده طبقه[2] مفاهیم با مشخصه معنایی مطرح گردد. در درک مفاهیم استعاری جدید از روش مقایسه و مطابقت مشخصه‌های معنایی استفاده می‌گردد. به عنوان مثال جمله «علی یخچال است» را در نظر بگیرید. در صورتی که این جمله را برای اولین بار می‌شنوید احتمالا کنجکاوید که بدانید چرا علی را یخچال گفته‌اند. مشخصه‌های معنایی یخچال که مشترک با انسان می‌تواند باشد «سردی» است. این به این معنی است که علی آدم سرد مزاجی است. در صورتی که این مفهوم استعاری در یک گونه زبانی به صورت مکرر استفاده شود واژه «یخچال» به عنوان نماینده مفاهیمی که سردمزاجی را نشان می‌دهند مطرح می‌شود و بنابر این درک آن نه از طریق تطبیق مشخصات معنایی بلکه از طریق دریافت مستقیم مفهوم استعاری انجام می‌گیرد.

عبدی(۱۳۹۰) برای نخستین بار روش‌های مختلف تشخیص و تفسیر زبان استعاری را در حوزه پردازش زبان فارسی به تفصیل شرح داده و به ارزیابی سامانه‌های مربوط به این حوزه در زبان‌های دیگر پرداخته است. در این قسمت با تاکید بر بخش استعاره، توضیحات وی در زمینه تفسیر با توجه به هدف اصلی تحقیق حاضر بررسی می‌شود.

تفسیر یا درک استعاره را نخستین بار مارتین[3](۱۹۹۰) تقریبا همزمان با تشخیص زبان استعاری توسط فاس، در سامانه‌ای به نام میداس[4] طراحی کرد. میداس با دریافت عبارت استعاره‌ای، پایگاه دانش خود را برای استعارهٔ متناظری جستجو می‌کرد که بی‌قاعدگی را توضیح دهد و در صورت نیافتن آن، سامانه به جستجوی استعاره در مفاهیم عام‌تر به کار خود ادامه می‌داد. در نهایت در صورت یافتن استعارهٔ عام‌تر، سامانه نگاشتی را برای زیرشاخه‌های آن یا استعارهٔ خاص‌تر ایجاد می‌کرد.

---

[1]- Gentner, D.
[2]- Typical Candidate of a semantic category
[3]- Martin,H.
[4]- MIDAS

تفسیر در دسته‌ای از دیدگاه‌ها یادگیری و استنتاج مشبه و مشبه‌به را در بر می‌گیرد؛ مانند روش«کارما»[۱] و پروژه«ای‌تی‌تی- متا»[۲](باردن[۳]، ۲۰۱۲). مجریان هر دو روش، چارچوبی استنتاجی طبق نظریهٔ استعارهٔ ادراکی برای استعاره تشکیل دادند. فرایند استنتاج آنها بر اساس دانشی است که از قبل در آن وجود دارد و بیشتر در حوزهٔ مشبه عمل می‌کند. سپس نتایج این استنتاج با استفاده از نمود نگاشتی ادراکی به حوزهٔ مشبه‌به منتقل می‌شود. اساس کار پروژه«ای‌تی‌تی-متا» توصیف استعاری حالت‌های ذهنی و همچنین استنتاج آنها با استفاده از منطق مرتبه اول صورت می‌گیرد. ورودی این روش، زبان طبیعی نیست، بلکه عبارات منطقی است که نمود چارچوب‌های کلامی است. «کارما» نیز به نوبه خود با اعمال و پیشامدهای انتزاعی سروکار دارد و متن تجزیه‌شده را به عنوان ورودی دریافت می‌کند.

وئال[۴] و هائو[۵] (۲۰۰۸) از «بازنمود دانش سیال برای تفسیر و تولید استعاره» با نام «نشانه‌های ناطق»[۶] یاد کردند. در روش آنها، با استفاده از مشخصات مشترک مفاهیم مبدا و مقصد و استفاده از دانش جهانی توسط «وردنت»[۷] و وب عمل درک استعاره انجام می‌شود. عمل یافتن مفاهیم مشترک مشترک با استفاده از«اسلیپ‌نت»[۸] که چارچوب ثبت دانش محیط خارج است انجام می‌پذیرد. در این این روش با استفاده از یک جمله مرتبط با ماهیت مفهوم مبدا و یک جمله مرتبط با ماهیت مفهوم مقصد سعی می‌شود تا با حذف و درج بعضی کلمات در جمله، از جمله مبدا به جمله مقصد برسیم.

شتووا[۹](۲۰۱۱)عمل درک استعاره را با استفاده از تغییر جمله‌بندی انجام داده است. او با استفاده از از پیکره ملی بریتانیا تلاش می‌کند تا یک جمله استعاری را به یک جمله غیر استعاری تبدیل کند. مثلا جمله شماره یک را به جمله شماره دو تبدیل کند:

(1) All of this stirred an unfathomable excitement in her
(2) This provoked an unfathomable excitement in her

وی با استفاده از مدل‌های احتمالاتی سعی می‌کند تا تمامی جمله‌بندی‌های مجاز برای یک جمله استعاری را بیابد و با استفاده از قواعد انتخابی یکی از آنها را به عنوان تفسیر جمله استعاری بیان کند.

1- KARMA
2- Att-Meta
3- Barden,J.A.
4- Veale,T.
5- Hao,Y.
6- Talking Points
7- WordNet
8- Slipnet
9- Shutova, E

۳- دادگان

۳-۱- واژه‌های دستوری

واژه‌های دستوری به طور پیش‌فرض فاقد مفهوم‌اند و تنها در بافت حضور می‌یابند و معنی پیدا می‌کنند. از آنجا که استعاره منوط به واژه‌های قاموسی و غیر دستوری است و واژه‌های دستوری مخل عملیات فرایند می‌شوند، در بررسی‌ها، فهرستی از این کلمات از واژه‌های زایای زبان فارسی استخراج و از پیکره مورد استفاده حذف شد. این فهرست بر اساس حوزه‌های صرفی، نحوی و معناشناختی جمع‌آوری شدند. در جدول شماره ۱، نمونه‌هایی از آن به همراه مثال‌هایی آورده شده است.

مقولهٔ نحوی	مثال	مقولهٔ نحوی	مثال
صفت اشاره	آن، همان ...	ضمیر پرسشی	آیا، چند، چرا...

جدول شماره ۱- نمونه‌هایی از کلمات دستوری

۳-۲- پیکره همشهری

پیکره همشهری (آل احمد و همکاران: ۲۰۰۹) پیکره استاندارد متنی به زبان فارسی است که به منظور استفاده در روش‌های بازیابی اطلاعات ایجاد شده است. این پیکره دارای ۱۶۶۷۷۴ مستند و ۶۵ موضوع است. این پیکره، دادگان آموزشی تحقیق حاضر را تشکیل می‌دهد.

۳-۳- استخراج صفات

صفات به این دلیل که مقوله اصلی مشبه و مشبه‌به‌اند، محصول تحقیق حاضر را تشکیل می‌دهند. سامانه سایه‌نما در نشان دادن خروجی نیاز به غربال‌گری دارد تا تنها صفات و امتیاز آنها را معرفی کند. در مراجعه دوباره به واژه‌های زایای زبان فارسی، تمامی کلمات با این برچسب استخراج و در غربال‌گری نهایی استفاده شد. به نمونه‌ای از این صفات در جدول شماره ۲ اشاره شده است.

درست‌آیین	باصفا	برتر	خشمگین	جوان	هیجان‌زده
دشنام‌گو	باسعادت	زیبا	کاری	سرد	خیال‌پرست

جدول شماره ۲- نمونه‌ای از صفات فاز غربال

۴- روش کار

۴-۱-تقطیع پیکره

در گام نخست از فرایند استخراج، تقطیع‌گر[1] زبان نشانه‌گذاری گسترش‌پذیر (اکس ام ال)[2] مد نظر قرار گرفت. زیرا مستندات موجود در پیکره همشهری، توسط زبان نشانه‌گذاری گسترش‌پذیر ذخیره شدند این مستندات دارای موضوع و متن موضوع است. بدین ترتیب نخست این مستندات تقطیع[3] و و سپس متن اصلی آنها استخراج شده است.

۴-۲- نرمال‌سازی متون و حذف کلمات دستوری

از آنجایی که متون موجود در پیکره همشهری دارای صورت‌های غیر استاندارد نویسه‌ها وکلمات است، قبل از آنکه بتوان از آنها در سامانه استفاده کرد باید ابتدا پیش‌پردازشی روی آنها انجام گیرد تا صورت‌های غیراستاندارد به شکل استاندارد تبدیل گردند. برای مثال، با استفاده از نرمال‌ساز، حروفی مانند «ی»، «ک» و همزه یکسان‌سازی شدند. در نهایت از آنجا که اعتقاد عمومی بر این است که در جملات زبان، واژه‌های دستوری در رساندن معنای متن تاثیر چندانی ندارند، حذف گردیدند. در پژوهش حاضر نیز از حضور آنها صرف نظر می‌شود و تنها واژه‌های قاموسی و غیر دستوری مجموعه دادگان پالایش‌شده سامانه حاضر را تشکیل می‌دهند.

۴-۳- استفاده از معیار اطلاعات مشترک نقطه‌ای به منظور یافتن بافت‌های مهم یک واژه

جهت اندازه‌گیری اهمیت هر یک از ویژگی‌های معنایی یک مفهوم از اطلاعات مشترک نقطه‌ای (چرچ[4] و هنکس[5]: ۱۹۸۹) استفاده شد. این معیار نشانگر میزان هم‌رخدادی دو رخداد، فارغ از میزان رخداد هر یک به صورت جداگانه است. اطلاعات مشترک نقطه‌ای دو رخداد x و y از دو متغیر تصادفی X و Y به صورت زیر محاسبه می‌گردد.

$$PMI(x,y) = \log_2 \frac{p(x,y)}{p(x) * p(y)}$$

هر چه میزان اطلاعات مشترک نقطه‌ای دو متغیر بیشتر باشد به این معنی است که تمایل به رخداد همزمان دو متغیر بیشتر است. مقدار صفر این معیار نشانگر شروع نقطه استقلال دو متغیر

---

[1]- parser
[2]- xml: extensible markup language
[3]- parse
[4]- Church, K.W.
[5]- Hanks, P.

است. همچنین مقدار منفی این معیار نشانگر عدم تمایل به هم‌رخدادی دو متغیر و یا خاصیت دفعی آنهاست.

این معیار جهت بررسی میزان تمایل هم‌رخدادی یک مفهوم در یک بافت به کار رفته است. در این بررسی، متغیر تصادفی اول، رخداد یک مفهوم و متغیر تصادفی دوم، رخداد یک بافت در نظر گرفته شده است. به عنوان مثال متغیر تصادفی رخداد واژه w2 در بافت _w1 را در نظر بگیرید. تمایل به هم‌رخدادی در صورتی بالاست که درصد بالایی از رخدادهای واژه w2 و بافت _w1 با هم انجام شده باشد. به این معنی که ممکن است هر کدام از این دو رخداد جداگانه به تعداد بالا حضور داشته باشند، اما رخداد همزمان این دو متغیر کم باشد. همچنین امکان دارد که این دو متغیر بسامد بسیار کم داشته باشند، اما در اکثر موارد هم‌رخداد باشد. معیار اطلاعات مشترک نقطه‌ای با حذف اثر میزان رخداد جداگانه هر یک از متغیرها، تنها نشانگر میزان تمایل به هم‌رخدادی را ارائه می‌دهد. در ادامه با مثال عددی این معیار بیشتر توضیح داده شده است. فرض کنیم که ۱۰۰۰ دنباله ۲تایی w1w2 در پیکره وجود دارد. می‌خواهیم میزان تمایل به هم‌رخدادی واژه w2 در بافت _w1 را به دست آوریم. فرض کنیم که w2 به تعداد ۵۰ بار ظاهر شده است. همچنین بافت _w1 با بسامد ۲۵ ظاهر شده است. بنابر این احتمال رخداد واژه w2 برابر ۰٫۰۵ و احتمال رخداد بافت _w1 برابر ۰٫۰۲۵خواهد بود. در صورتی که w2 در _w1 در ۵ دنباله ۲تایی هم‌رخداد باشند، اطلاعات دوسویه مشترک آنها از طریق فرمول زیر به دست می‌آید.

$$PMI(w2, w1_) = \log_2 \frac{p(w2, w1_)}{p(w2) * p(w1_)} = \log_2 \frac{\frac{5}{1000}}{\frac{50}{1000} * \frac{25}{1000}} = 2$$

در این فرمول p(w2,w1_) برابر با ۰٫۰۰۵، p(w2) برابر با ۰٫۰۵ و p(w1_) برابر با ۰٫۰۲۵ است. بنابر این مقدار اطلاعات مشترک نقطه‌ای برابر با ۲ می‌گردد. مثبت بودن این معیار نشانگر تمایل رخداد w2 در بافت _w1 است.

در صورتی‌که فرض کنیم واژه w2 به تعداد ۸۰۰ بار در ۱۰۰۰ دنباله دوتایی رخداده باشد و بافت _w1 به تعداد ۴۰۰ بار در ۱۰۰۰ دنباله دوتایی به کار رفته باشد و همچنین تعداد هم‌رخدادی این واژه در بافت ذکر شده برابر با ۲۰ باشد، میزان اطلاعات مشترک نقطه‌ای به طریق مشابه به دست خواهد آمد.

$$PMI(w2, w1_) = \log_2 \frac{p(w2, w1_)}{p(w2) * p(w1_)} = \log_2 \frac{\frac{20}{1000}}{\frac{800}{1000} * \frac{400}{1000}} = -4$$

مقدار منفی اطلاعات مشترک نقطه‌ای نشان می‌دهد که علی رغم اینکه واژه و بافت در حالت دوم نسبت به حالت اول به تعداد بیشتری جداگانه رخ می‌دهند و حتی تعداد رخداد همزمان آنها بیشتر است اما از آنجا که w2 تمایل بیشتری به هم‌رخدادی در بافت‌های دیگر دارد تا رخ دادن در بافت w1_ میزان اطلاعات مشترک نقطه‌ای آنها منفی شده است.

۴-۴- تولید سه‌تایی‌ها[1] از پیکره

دنباله واژه‌های موجود در پیکره را به صورت Corpus={w1, w2, …, wn}در نظر بگیرید. با استفاده از برنامه ساده سه‌تایی‌هایی به شکل {w1, w2, w3}, {w2, w3, w4}, {w3,w4,w5}, …,{wn-2, wn-1, wn}از پیکره استخراج و ذخیره شدند. به دلیل حجم بالای پیکره از چارچوب پردازش توزیع شده هدوپ[2] و مدل برنامه نویسی نگاشت/کاهش[3] برای تولید این سه‌تایی‌ها از شبکه کامپیوتری استفاده شد. سپس هر یک از کلمات کدگذاری عددی و سه‌تایی‌ها به صورت کدگذاری عددی در پایگاه داده ذخیره شدند. در هر سه‌تایی واژه‌های اول و سوم بافت زبانی را تشکیل می‌دهند که واژه دوم در آن رخ داده است. برای تعیین بافت، روش‌های مختلفی وجود دارد. در برخی از روش‌ها مانند مدل‌های زبانی «چندتایی یا چندنگاشتی»[4] فرض بر این است که احتمال رخداد هر واژه تنها به واژه‌های قبلی خود وابسته است و مستقل از واژه‌های بعدی است. اما این فرض در اینجا نمی‌تواند صحیح باشد چرا که هدف از این بررسی شناخت مفهوم مشترک دو واژه است. سه‌تایی‌های «اشتهای گرگ سنگدل» و «خشونت مرد سنگدل» را در نظر بگیرید و فرض کنید که می‌خواهیم مفهوم مشترک بین دو واژه مرد و گرگ را در یک گزاره استعاری بیابیم. در این سه‌تایی واژه‌های «مرد» و «گرگ» در میانه بافت قرار گرفته‌اند و واژه پس از این واژه‌ها می‌تواند به ما در شناخت مفهوم مشترک کمک کند. همچنین واژه‌های رخداده پیش از واژه‌های «مرد» و «گرگ» نیز می‌تواند مفهوم مشترک دو واژه باشند و به ما در درک معنای استعاری کمک کنند. بنابر این در این بررسی هم از بافت پیشین و هم از بافت پسین برای یافتن مفهوم مشترک دو واژه استفاده شد.

۴-۵- انتخاب بافت‌های مشترک به عنوان مفهوم معنایی مشترک دو واژه

در مرحله قبل بافت‌های مهم هر یک از دو واژه را با استفاده از اطلاعات مشترک نقطه‌ای یافتیم. در این مرحله باید بافت‌هایی از هر دو واژه را بیابیم که مفهوم معنایی مشترک دو واژه محسوب شوند و ما را در درک بهتر این استعاره یاری کنند. بافتی که به عنوان مفهوم مشترک دو واژه در درک

---

[1]-trigrams
[2]- Apache Hadoop
[3]- MapReduce
[4]- n-gram

استعاری وجود دارد، الزاما یکی از با اهمیت‌ترین ویژگی‌های معنایی واژه مبدا است. بنابر این برای انتخاب بافت‌های معنایی از مجموع خطی وزن‌دار بافت‌ها استفاده می‌کنیم. در فرمول زیر شیوه محاسبه امتیاز بافت‌های مشترک آمده است.

$$\text{Score (context, source, target)} = C * \text{Score (context, source)} + (1- C) * \text{score (context, target)}$$

در این فرمول، C باید بین ۰ و ۱ و بیشتر نزدیک به ۱ باشد تا سهم امتیاز بافت‌های واژه مبدا بیشتر از واژه مقصد شود. در این آزمایش‌ها C را برابر ۰/۸در نظر گرفته‌ایم.

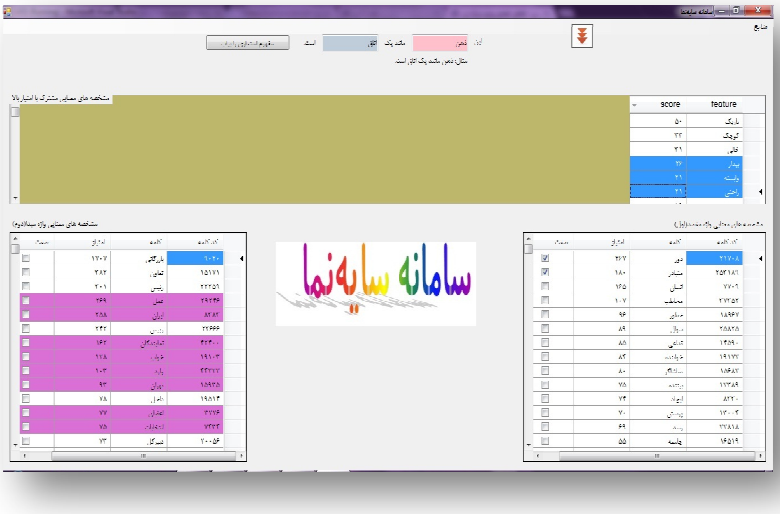
۵- نتایج

در جدول زیر مفهوم مشترک تعدادی از جفت‌واژه‌های تولید شده توسط سامانه سایه‌نما به همراه امتیاز هر کدام مشاهده می‌شود.

واژه مقصد – واژه مبدا												
ذهن – اتاق		شادی – جرقه		ذهن- ساعت		مرد – روباه		مرد – گرگ		مرد – خرس		خانه – بهشت	
تاریک	۵۰	زودگذر	۴۲	ویژه	۴۸	قاتل	۴۶	مهاجم	۲۴	غارنشین	۵۷	امن	۱۴
کوچک	۳۳	روشن	۱۶	جوان	۴۰	پیر	۴۱	پیر	۲۳	خشمگین	۲۰	خلوت	۹
خالی	۳۱	کوتاه	۱۶	کاری	۳۱	بیچاره	۲۹	خشمگین	۲۰	قوی	۵	زیبا	۱

جدول شماره ۳- مفهوم مشترک تعدادی از جفت‌واژه‌های تولیدشده توسط سامانه

برای مشاهده سامانه تولید شده در این بررسی، نمونه‌ای از آزمایش آن در شکل شماره ۱ قابل مشاهده است.



شکل۱-شمای کلی سامانه سایه‌نما

۶- بحث

با توجه به اینکه پیکره دستی از استعاره‌ها در زبان فارسی وجود ندارد، امکان ارزیابی این سامانه در حجم وسیع امکان‌پذیر نیست. برای ارزیابی این سامانه ابتدا باید پیکره‌ای از استعاره‌ها در زبان فارسی تولید شود تا در ارزیابی روش‌های درک استعاره به کار گرفته شود. در ارزیابی‌هایی که توسط نگارندگان به صورت موردی انجام شد، این سامانه به تغییراتی برای دستیابی به جواب‌های بهینه‌تر نیاز دارد که در ادامه به آن پرداخته شده است.

برای بهبود کیفیت سامانه، بهینه‌سازی باید انجام گردد:

- استفاده از وردنت فارسی برای یافتن مفاهیم مادر یک مفهوم در سلسله مراتب مفاهیم و استفاده از صفت‌های مادر یک مفهوم در عمل تداخل باعث می‌شود تا تداخل بهتر انجام گردد. چرا که در بسیاری از موارد مفهوم مبدا به عنوان پیش‌نمونه‌ای[1] از مفهوم مادر مطرح می‌گردد.

– استفاده از پیکره جامع‌تر: پیکره همشهری علی رغم حجم بالای آن بیشتر شامل اخبار سیاسی است که با توجه به ماهیت اخبار سیاسی در کشور از مشخصه‌های خاص برخوردار است و جامعیت یک پیکره از نظر پوشش مفاهیم را دارا نیست. همچنین از استفاده خلاقانه گسترده بی‌بهره است. به عنوان مثال در این پیکره واژه آمریکا در بسیاری از مفاهیم وارد می‌شود، در صورتی که در گفتار عادی یک فارسی زبان اینگونه نیست و این برگرفته از فرهنگ سیاسی است.

- استفاده از مدل‌های تحلیل معنایی پنهان می‌تواند به عنوان راه کاری جهت کاهش بعد و همچنین افزایش یادآوری سامانه استفاده گردد. در این مدل‌ها کلمه‌های هم‌معنی در یک خوشه قرار می‌گیرد و خوشه به عنوان مشخصه معنایی یک مفهوم در نظر گرفته می‌شود.

- استفاده از مدل‌های وزن‌دهی پیچیده‌تر برای مشخصه ضروری به نظر می‌رسد. در پروژه کنونی از بسامد رخداد مشخصه معنایی در کنار مفهوم مبدا یا مقصد به عنوان وزن آن مشخصه استفاده شده است. این وزن‌دهی می‌تواند به وزن‌دهی‌های پیچیده‌تر مانند tf-idf در حوزه چندنگاشتی تبدیل شود. در این روش هر چندنگاشت به عنوان یک سند محسوب می‌گردد.

- استفاده از بافت در تشخیص بهترین مشخصه‌های مشترک برای درک استعاره می‌تواند بسیارمفید باشد. به عنوان مثال در بافت زمستانی، اگر یک فرد را گرگ بنامیم ممکن است ناشی از صبر او در سرما باشد اما در بافت غیر زمستانی و غیر کوهستانی این مشخصه به بی‌رحمی تعبیر خواهد شد.

[1]- Proto-type

منابع

- اسلامی، محرم، مسعود شریفی، صدیقه علیزاده و طاهره زندی(۱۳۸۳).«واژگان زایای زبان فارسی»، *مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه*، صص ۶-۱۱.

- عبدی قویدل، هادی، محمد بحرانی و بهرام وزیرنژاد(۱۳۹۱). «بررسی روش‌های تشخیص خودکار و رفع ابهام معنایی مفاهیم استعاری و دارای مجاز». دومین هم‌اندیشی زبان‌شناسی رایانشی، صص ۳۳-۴۴.

- AleAhmad, A. and others(July 2009). "Hamshahri: A Standard Persian Text Collection", *Journal of Knowledge-Based Systems*, Vol. 22 No.5, p.382-387, Elsevier,.

- Barnden, J.A. (2012). ATT-Meta Project Databank. http:// www.cs.bham.ac.uk//˜jab/ATT-Meta/Databank.

- Church, Kenneth Ward and P. Hanks(1989)." Word Association Norms, Mutual Information, and Lexicography",*Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*.

- Fass, D. (1991)." met*: A Method for Discriminating Metonymy and Metaphor by Computer". *Computational Linguistics*, 17(1):49–90.

- Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7(2),155-170.

- Islam, A. and D. Inkpen(2008). "Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity". *ACM Trans. Knowl. Discov.* Data 2, 2 (Jul. 2008), 1–2.

- Islam, A. and D.Inkpen(2006)." Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words", *in Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, pp. 1033–1038.

- Martin, H. (1990). A Computational Model of Metaphor Interpretation. *Academic Press Professional, Inc.*, San Diego, CA, USA.

-McCallum-Bayliss, H.(2011) Metaphor Program Proposer. *INTELLIGENCE ADVANCED RESEARCH PROJECTS ACTIVITY (IARPA)* http://www.iarpa.gov

- Shutova, E. (2011). *Computational Approaches to Figurative Language.PhD* Thesis, Computer Laboratory, University of Cambridge, UK.
-Veale, T. and Y.Hao(2008). A Fuid Knowledge Representation for Understanding and Generating Creative Metaphors. *In Proceedings of the 22nd International Conference on Computational Linguistics* (Coling 2008), pages 945–952, Manchester, UK.