استخراج ترکیب‌های وصفی استعاری، مبتنی بر پیکره و با استفاده از مدل جاسازی واژه‌ها

حمزه مطهری خوانساری، مهرنوش شمس فرد
دانشگاه شهید بهشتی، h_motahari@sbu.ac.ir
دانشگاه شهید بهشتی، m-shams@sbu.ac.ir

چکیده

استعاره را می‌توان سخن در دامنه‌ای با کمک از واژه‌هایی که به دامنه دیگری تعلق دارند دانست. پردازش استعاره‌های زبانی، یکی از حوزه‌های دشوار در پردازش زبان طبیعی می‌باشد. یکی از چالش‌های ابتدایی در این حوزه نبود داده استعاری مناسب، به ویژه برای زبان فارسی است. این مقاله در پی آن است که با کمک یک پیکره زبانی بزرگ، به صورت حداکثری انواع گونه‌های زبانی فارسی را به دست آورد و سپس بر اساس آن مدل جاسازی واژه‌هایی مناسب و پوشا تولید نماید. در ادامه با کمک این پیکره و مدل زبانی به دست آمده، بر مبنای این فرض که به کار بردن استعاره در زبان بشر، نوعی ناهنجاری به نسبت روش معمول سخن وی محسوب می‌شود، روشی برای تشخیص و استخراج ترکیب‌های وصفی استعاری تک‌واژه‌ای ارائه دهد و در نهایت یک مجموعه داده استعاری فراهم آورد.

واژه‌های کلیدی

استعاره، جاسازی واژه‌ها، پیکره، ترکیب‌های وصفی، وبلاگ

۱- مقدمه

یکی از حوزه‌های کار در پردازش و نیز فهم زبان طبیعی، پردازش استعاره‌ها در زبان می‌باشد. استعاره‌ها خود زیرمجموعه‌ای از زبان کنایی[۱] می‌باشند(Colston, 2015; Croft & Cruse, 2004). به طور خلاصه استعاره را می‌توان صحبت درباره مفهوم یا حوزه‌ای، با کمک کلماتی که کاملا به حوزه دیگری تعلق دارند، دانست (Jurafsky & Martin, 2008; Kövecses & Benczes, 2010). به عنوان مثال ترکیبات مشخص شده در جملات زیر استعاری می‌باشند:

- جمعیت *غرق* سخنرانی او *شده بودند*. (استفاده از واژه‌های دامنه دریانوردی یا شنا، در دامنه سخنرانی)
- رفع این مشکل اقتصادی تنها با یک *جراحی* وسیع قابل انجام است. (استفاده از واژه‌های دامنه پزشکی برای سخن در دامنه اقتصاد)
- وزارت کشور با مشورت گروه‌های کارشناسی متعدد در نهایت به تصمیم *پخته‌ای* رسید. (استفاده از دامنه آشپزی برای سخن در دامنه سیاست)

استعاره‌ها یکی از پیچیده‌ترین بخش‌های زبان بشر می‌باشند (Lakoff & Johnson, 2008) که در پردازش و فهم زبان طبیعی، به ویژه در زبان فارسی، بسیار جای کار دارند. پردازش استعاره‌ها به دو حوزه کلی تشخیص استعاره[۲] و تفسیر استعاره[۳] قابل‌تقسیم است (Shutova, 2016) که مجموع کارهای انجام‌شده در حوزه تشخیص به مراتب بیشتر است.

در تشخیص استعاره سطوح مختلفی مطرح است که عبارت‌اند از (Shutova, 2016):

- تشخیص جمله استعاری، بدون مشخص کردن دقیق جای استعاره(ها) در جمله.
- تشخیص دقیق عبارت استعاری در جمله، که خود به دو بخش تقسیم می‌شود:
  - عبارت تک‌واژه‌ای
  - عبارت چندواژه‌ای
- تشخیص استعاره فراجمله‌ای[۴]

این مقاله به تشخیص استعاره‌های تک‌واژه‌ای، در ترکیب‌های وصفی جملات، با استفاده از پیکره زبانی بزرگ و مدل جاسازی واژه‌ها[۵] پرداخته است. منظور از استعاره‌های تک‌واژه‌ای در ترکیب‌های وصفی، تشخیص ترکیب‌هایی وصفی و استعاری است، که شامل دو واژه موصوف و صفت در نزدیک هم، و به صورت تک‌واژه‌ای می‌باشند (مثل «لبخند سرد»).

۲- کارهای مرتبط

سابقه اولین کارها در حوزه تشخیص استعاره به اوایل دهه ۹۰ میلادی باز می‌گردد (Fass, 1991). در حوزه تشخیص استعاره فنون مختلفی به کار بسته شده است که البته بعضا، نه به تنهایی، بلکه به صورت ترکیبی مورد بهره‌برداری قرار گرفته‌اند.

یکی از مهم‌ترین و ابتدایی‌ترین روش‌ها برای تشخیص استعاره استفاده از نقض ترجیحات گزینشی می‌باشد. که می‌توان گفت بر اساس این ایده که تشخیص استعاره، تشخیص یک ناهنجاری در مقایسه با حالت رایج زبان در یک جمله است (Su, Huang, & Chen, 2017) شکل گرفته است. کارهای متعددی از این فن بهره برده‌اند (Fass, 1991; Gershman & Dyer, 2014; Haagsma & Bjerva, 2016; Jia & Yu, 2008; Mason,

---

[۱] یا زبان تمثیلی، یا زبان نمادین، Figurative Language

[۲] Metaphor Detection، در برخی ادبیات Metaphor Identification هم استفاده می‌شود.

[۳] Metaphor Interpretation

[۴] Extended Metaphor

[۵] Word Embeddings

2004; Neuman et al., 2013; Shutova & Sun, 2013). می‌توان گفت روش پیشنهادی این مقاله نیز، بر این اساس شکل گرفته است.

فن دیگری که می‌تواند مورد استفاده قرار گیرد استفاده از قواعد نحوی است (Krishnakumaran & Zhu, 2007). مثلا در یک قاعده ساده، هرگاه در یک جمله با ساختار Is_A، مسندالیه به صورت اسم نکره بیاید، به صورت بالقوه احتمال رخداد استعاره وجود دارد. به عنوان نمونه:

He is *a Gandhi*.

بدیهی است که در اینجا یک فرد نمی‌تواند دقیقا گاندی باشد. همچنین گاندی نیز یک شخص حقیقی است و معنا ندارد که به صورت نکره مورد استفاده قرار گیرد، و در واقع این ناهنجاری، نشان از وقوع یک حالت کنایی/استعاری می‌باشد. طبعا این روش، برای تشخیص برخی موارد سرراست و قاعده‌دار جواب‌گو است، اما در حالت کلی و به تنهایی کارساز نیست.

یک روش دیگر مورد استفاده، بهره بردن از روابطی مثل شمول و نیز چندمعنایی در وردنت (Miller, 1995) می‌باشد (Krishnakumaran & Zhu, 2007; Wilks, Dalton, Allen, & Galescu, 2013). مثلا جمله‌های زیر را در نظر بگیرید:

- خواهرشوهرهای او افعی هستند.
- خورشید خود یک ستاره است.

این جملات هر دو در حال بیان یک رابطه Is_A می‌باشند. اما با مراجعه به دادگان واژگانی[۶] (چون فارس‌نت Shamsfard, Hesabi, et al., 2010)، مشهود است که «خورشید» و «ستاره»، دارای سلسله‌مراتب مشترکی می‌باشند، در حالی که «خواهرشوهر» و «افعی» چنین وضعی ندارند. بنابراین چنین تفاوت‌ها و عدم انطباق‌هایی، نامزدهای مناسبی برای رخداد استعاره محسوب می‌شوند.

همچنین از روش‌های آماری از نوع دسته‌بندی[۷] در (Dunn, 2013; Gedigian, Bryant, Narayanan, & Ciric, 2006; Klebanov, Leong, & Flor, 2015; Turney, Neuman, Assaf, & Cohen, 2011) و نیز خوشه‌بندی[۸] در (Birke & Sarkar, 2006; Pramanick & Mitra, 2018; Shutova, Sun, & Korhonen, 2010) بهره برده شده است.

در استفاده از روش‌های دسته‌بندی معمولا مجموعه داده مورد بررسی کوچک می‌باشد و هر کس از زاویه‌ای به موضوع ورود کرده است و بر اساس داده‌های برچسب‌خورده بعضا متفاوت، و با استخراج ویژگی‌ها و روش‌های مختلف، اقدام به دسته‌بندی شده است.

همچنین می‌توان گفت در روش‌های مبتنی بر خوشه‌بندی نیز، بعضا خوشه‌بندی چندان به صورت خالص (و بدون نظارت) انجام نشده است، بلکه داده اولیه‌ای به صورت برچسب‌خورده تهیه شده است و بر اساس آن و با کمک خوشه‌بندی کار گسترش یافته و به انجام رسیده است. علاوه بر این در سال‌های اخیر موج استفاده از فنون یادگیری عمیق، به این حوزه نیز رسیده است (Bizzoni & Ghanimifard, 2018; Swarnkar & Singh, 2018; Wu et al., 2018).

در این میان، اصولا یکی از نقص‌های مهم در حوزه پردازش استعاره‌ها، کمبود مجموع داده‌های مناسب، به اندازه کافی بزرگ و استاندارد می‌باشد. مشهورترین پیکره موجود، پیکره VU Amsterdam (Steen, Dorst, Herrmann, Kaal, & Krennmayr, 2010) است که حاوی ۴۶۶۸ جمله انگلیسی است که به صورت دستی برچسب خورده‌اند. این عدم وجود مجموعه داده مناسب، هم خود پژوهش در این حوزه را دشوار نموده است و هم مقایسه و ارزیابی کارها را دور از دسترس می‌نماید. این پژوهش گامی است که در این راستا و برای زبان فارسی برداشته شده است.

۳- روش پیشنهادی

همان‌طور که دربخش‌های پیشین گفته شد، هدف این پژوهش تشخیص و استخراج ترکیب‌های وصفی استعاری و تهیه مجموعه داده‌ای بر این اساس می‌باشد. در ادامه دیده خواهد شد، در راه رسیدن به این هدف، نیاز به استفاده از ابزارهای پایه پردازش زبان طبیعی در زبان فارسی، که در دسترس است، می‌باشد. البته می‌توان گفت عموما ابزارهای استفاده شده در این پژوهش، با دقت و کارایی مورد انتظار فاصله داشته‌اند. بر همین اساس و برای کمک به پوشش خطاهای ابزار، روش پیشنهادی این مقاله، به طور عمده بر پایه وجود یک پیکره بزرگ که برای انواع گونه‌های زبان فارسی به اندازه کافی پوشا باشد نهاده شده است. تا بلکه بتوان با داده بسیار زیاد، اثر خطاهای انسانی و ابزارها تا حد ممکن کاهش یابد. از آنجا که با بررسی‌های انجام‌شده، پیکره مناسب این پژوهش یافت نشد، اقدام به تهیه پیکره‌ای متنی و بزرگ برای این کار شد.

در ادامه چگونگی تهیه پیکره و مشخصات آن، چگونگی تهیه ترکیب‌های وصفی، و سپس روش پیشنهادی بیان می‌گردد.

۳-۱- تهیه پیکره

یکی از منابع مهم متنی عظیم و در دسترس زبان فارسی، که گونه‌های مختلفی از این زبان را پوشش داده است، وبلاگ‌ها می‌باشند. این داده‌ها، که بر خلاف برخی داده‌های امروزی شبکه‌های اجتماعی، با آسانی بیشتری قابل خزش و دریافت می‌باشند، به عنوان منبع تهیه پیکره در نظر گرفته شده است. یک امتیاز مهم وبلاگ‌ها، گستردگی و تنوع نویسنده‌ها، متن‌ها، و قدمت و سابقه کاملا طولانی آنها، از باب پوشش زبان برای مدتی طولانی، می‌باشد. گرچه این گزاره شنیدنی است که وبلاگ‌ها در رقابت با شبکه‌های اجتماعی امروزی، میدان را واگذار کرده‌اند و اقبال کمتری یافته‌اند.

بیشتر وبلاگ‌ها، در محیط فارسی‌زبانان، توسط سرویس‌دهندگان متعددی به رایگان میزبانی می‌شوند. در این میان سرویس‌دهنده بلاگفا احتمالا پرمشتری‌ترین میزبان وبلاگ‌های فارسی است. برای همین وبلاگ‌های این سرویس‌دهنده، به عنوان منبع تهیه پیکره این پژوهش، موسوم به پیکره hmBlogfa، در نظر گرفته شد.

---

[۶] Lexical Database

[۷] Classification

[۸] Clustering

برای خزش و جمع‌آوری پست‌های وبلاگ‌ها، در گام نخست، نیاز به فهرستی از آنان بود. چنین فهرستی به صورت عمومی در دسترس نمی‌باشد. برای همین یک فهرست اولیه با حدود ۴۰هزار نشانی وبلاگ، از دو منبع زیر تهیه شد:

- خزش پیوندهای وبلاگ‌های (بروزشده و ...) قابل‌استخراج از صفحه اصلی بلاگفا
- استخراج نشانی، از یک مجموعه داده نمونه، که توسط یکی از مسئولین موتور جستجوی یوز ("Yooz search engine," n.d.)، در اختیار این پژوهش قرار گرفت.

الگوریتم خزش به صورت خلاصه به شکل زیر بوده است:

- یک نشانی به صورت نیمه‌تصادفی[۹] از بین نشانی وبلاگ‌های خزش‌نشده انتخاب می‌گردد و صفحه اصلی آن وبلاگ خزش می‌شود.
- از صفحه اصلی دریافتی، پیوندهای احتمالی این وبلاگ با وبلاگ‌های دیگر کشف می‌شوند و بدین ترتیب نشانی‌های جدید، برای خزش بعدی ثبت می‌شوند و در صف قرار می‌گیرند.
- از صفحه اصلی، شماره آخرین پست کاربر استخراج می‌گردد.
- به صورت تصادفی، ۱۰۰ پست در بازه پست اول و پست آخر وبلاگ خزش می‌شود. در این میان وبلاگ‌هایی که مجموعا کمتر از ۱۰۲ پست داشته باشند، عملا تمام پست‌های‌شان[۱۰] خزش می‌شوند. البته ممکن است برخی پست‌های انتخاب‌شده برای خزش، قبلا توسط نویسنده حذف شده باشند.

خزش بلاگفا حدود دو ماه به طول انجامید. در این خزش حدود ۴۰۰هزار نشانی وبلاگ با موفقیت خزش شد و قریب ۱۹میلیون پست دریافت شد. روی این ۱۹میلیون پست پیش‌پردازش‌هایی به صورت زیر انجام شده است:

- تبدیل نویسه‌های نااستاندارد به نویسه‌های استاندارد فارسی (مثل تبدیل «ی» و «ک» عربی)[۱۱].
- حذف اعداد و جایگزینی تمام آنها با عدد ۱۳۹۸.
- حذف جملات تماما غیرفارسی.
- حذف پست‌هایی که بیش از آنکه فارسی باشند به زبان دیگری هستند (مانند پست‌های انگلیسی، عربی، ترکی و ...)؛ با کمک یک تشخیص‌دهنده زبان که در این پژوهش توسعه یافته بود.
- اصلاح واژه‌های با نویسه‌های تکراری بیش از ۲ بار و پشت‌سرهم، که استفاده از آنها در وبلاگ‌نویسی رایج است (مانند جابجایی «خووووووب» با «خوب»).
- حذف پست‌های خالی از متن (مانند پست‌هایی که فقط حاوی تصویر بوده‌اند) و نیز پست‌های تقریبا خالی از متن.
- حذف پست‌های رمزشده[۱۲].

حاصل کار، پس از این پیش‌پردازش‌ها، حدود ۱۶٫۵ میلیون پست، شامل حدود ۴۰۰میلیون جمله و پنج میلیارد واژه می‌باشد.

پس از این مرحله، اقدام به تهیه مدل جاسازی واژه‌ها بر پایه این پیکره گردید. برای این کار از ابزار genism (Řehůřek & Sojka, 2010)، با پارامترهایی به صورت اندازه بُعد ۴۰۰، طول پنجره ۵ و حالت skip-gram (Mikolov, Chen, Corrado, & Dean, 2013) بهره برده شده است.

با توجه به گستردگی و تنوع این پیکره، واژه‌های بسیار نادری بابت سلیقه یا زبان خاص کاربران، رسم‌الخط‌های مختلف آنان، و نیز اشتباهات متعدد املایی و نگارشی یافت می‌شدند که بیش از این که کمک به مدل زبانی کنند، در واقع منبع افزایش خطا و بی‌نظمی بودند. بر این اساس و برای کاهش این مساله، برابر بررسی‌های صورت‌گرفته به نظر رسید که واژه‌های با بسامد بیشتر از دو هزار بار، مبنای تهیه مدل قرار گیرند. شایان ذکر است که با حذف واژه‌های با بسامد کمتر از این میزان، همچنان بیش از ۹۷٪ پیکره برای آموزش مدل حفظ شد.

همچنین به نظر رسید که برای هدف این پژوهش مناسب‌تر آن است که از پیکره قطعه‌بندی و ریشه‌یابی‌شده استفاده گردد[۱۳]. با توجه به حجم بالای پیکره و محدودیت‌های زمانی و پردازشی، امکان اینکه کل پیکره مستقلا به ابزار داده شود تا قطعه‌بندی و ریشه‌یابی شود فراهم نبود. برای همین فهرست واژه‌های پیکره مستقلا به ابزار استپ‌وان (Shamsfard, Jafari, & Ilbeygi, 2010) داده شد و سپس این خروجی، برای تبدیل و جایگزینی موارد اصلاح‌شده روی پیکره، به کار گرفته شد. البته خود این فرآیند قطعه‌بندی و ریشه‌یابی، بابت خطاهای ابزار، موجب انتشار برخی اشتباهات در پیکره گشته است. البته با برخی ویرایش‌های مختصر دستی، سعی شد حتی‌الامکان دامنه انتشار این خطاها کاهش یابد.

۳-۲- تهیه ترکیب‌های وصفی

گام بعدی پس از تهیه پیکره، تهیه فهرستی از ترکیب‌های وصفی می‌باشد. منظور از ترکیب‌های وصفی، ترکیب یک اسم با یک صفت می‌باشد (مانند «پسر خوب») و از حالت‌های پیچیده‌تر، مانند ترکیب اسم با چند صفت (مانند «پسر خوب قدبلند») چشم‌پوشی شده است. همچنین با توجه به

---

[۹] دلیل تعبیر نیمه‌تصادفی آن است که با توجه به مجموعه داده اولیه نشانی وبلاگ‌ها، و نیز بند بعد الگوریتم، نمی‌توان فرآیند انتخاب وبلاگ‌ها برای خزش را فرآیندی کاملا تصادفی نامید.

[۱۰] به جز دو پست اول و آخر. از آنجا که احتمال دارد پست‌های ابتدایی و بعضا انتهایی وبلاگ‌ها (وبلاگ‌های غیرفعال) واجد نوشته‌هایی با مضمون مشابه باشند، برای جلوگیری از سوگیری، این پست‌ها در نظر گرفته نشده‌اند.

[۱۱] با توجه به محدود بودن استفاده از نیم‌فاصله در پست‌های وبلاگ‌ها، برای رسیدن به یک‌دستی در پیکره، این نویسه با فاصله جایگزین شد.

[۱۲] یک امکان در بلاگفا، که پست‌ها می‌توانند رمزدار باشند و فقط توسط افرادی که رمز را بدانند قابل‌خواندن است.

[۱۳] توضیح بیشتر در بخش ۴ آمده است.

خطای قابل‌توجه ابزار، ترکیب‌های وصفی که بین صفت و موصوف از دید تجزیه‌گر وابستگی فاصله افتاده باشد نیز حذف گردیدند[۱۴].

برای این منظور اقدام به جداسازی ۴۰۲۷۰۴۲ جمله یکتا با طول بین ۱۶ تا ۱۲۸ نویسه از ابتدای پیکره hmBlogfa گردید. این مقدار با توجه به طول جمله متعارف، انتخاب گردیده است. از آنجایی که در بین وبلاگ‌نویسان بعضا چندان اهتمامی به رعایت قواعد نگارشی وجود ندارد، برخی جملات بسیار طولانی می‌باشند و در واقع از لحاظ ادبی و نگارشی یک جمله محسوب نمی‌شوند. همچنین جملات خیلی کوتاه نیز احتمالا نامزد مناسبی برای استخراج ترکیب‌های وصفی نیستند.

سپس این جمله‌ها به ابزار تجزیه‌گر وابستگی استنفورد (Qi, Dozat, Zhang, & Manning, 2018) داده شد تا جملات از نظر وابستگی تجزیه شوند. سپس از این خروجی، ترکیب‌های Noun_Adj استخراج گردید؛ که حاصل استخراج ۲۵۹۱۱۱۴ ترکیب وصفی نایکتا می‌باشد؛ که پس از یکتاسازی معادل ۹۱۷۴۱۳ ترکیب وصفی یکتا به دست آمد. البته طبیعتا بابت ضعف املایی و نگارشی جملات ورودی، و نیز خطای ابزار، تمام ترکیبات تولیدشده صحیح نمی‌باشند و اشتباهات متعددی در میان آنان یافت می‌شود.

در هر حال، با آماده شدن این ترکیب‌ها، داده مناسب برای اجرای روش پیشنهادی این مقاله، برای یافتن ترکیب‌های وصفی استعاری فراهم آمد که در بخش بعد توضیح داده می‌شود.

۳-۳- استخراج ترکیب‌های وصفی استعاری

اساس کار در روش پیشنهادی این مقاله برای استخراج ترکیب‌های وصفی، پیدا کردن راهی برای پی بردن به این پرسش است که آیا یک ترکیب وصفی، در زبان مرسوم است و توسط گویش‌وران تولید می‌شود؛ یا به نوعی ناهنجاری و نامعمول به حساب می‌آید. برای این کار از مدل جاسازی واژه‌ها به اضافه دوگرم[۱۵] مرتب‌شده بر اساس بسامد رخداد در پیکره، استفاده شده است.

شرح الگوریتم روش به صورت زیر است:

- یک ترکیب اسم-صفت انتخاب می‌شود.
- در مدل جاسازی واژه‌های به دست آمده، و با کمک gensim، ۲۰۰ واژه نزدیک به اسم و صفت، به صورت جداگانه، بازیابی می‌شوند. از آنجا که مدل جاسازی واژه‌ها بر اساس پیکره ریشه‌یابی‌شده تهیه شده است، ممکن است اسم یا صفت برخی ترکیب‌ها در مدل یافت نشوند. در این موارد چنین ترکیباتی در نظر گرفته نمی‌شوند و ترکیب بعدی اسم-صفت انتخاب می‌گردد.
- ۴۰۰ ترکیب جدید تولید می‌شوند که حاصل ترکیب اسم با ۲۰۰واژه نزدیک به صفت، و صفت با ۲۰۰واژه نزدیک به اسم می‌باشند.
- این ۴۰۰ ترکیب جدید به اضافه خود ترکیب، در فهرست دوگرم‌ها که از پررخداد به کم‌رخداد مرتب شده است جستجو می‌شوند و رتبه آنان کشف می‌گردد. در مواردی که ترکیب اصلا در دوگرم‌ها نباشد، مقدار آخرین رتبه فهرست دوگرم‌ها + ۱ (بالاترین عدد، به معنای کمترین رخداد) تخصیص می‌یابد.
- رتبه میانگین، بر اساس میانگین‌گیری روی کل رتبه‌های به دست آمده از ۴۰۰ ترکیب جدید و ۴۰برابر رتبه خود ترکیب، به دست می‌آید. دلیل ضریب ۴۰برابر داده‌شده به خود ترکیب، آن است که اثر رایج بودن/نبودن خود ترکیب در پیکره، حفظ شود (بدین ترتیب مجموع رتبه‌ها برای میان‌گیری تقسیم بر ۴۴۰ می‌شوند).

بر این اساس، ترکیب‌های با میانگین عددی پایین، احتمالا ترکیب‌های رایج در زبان هستند و استعاری محسوب نمی‌شوند و ترکیب‌های با میانگین عددی بالا، ترکیب‌های نادری هستند که ناهنجاری محسوب می‌شوند و نامزد استعاری بودن می‌باشند. همان‌طور که در بخش ۵ آمده است می‌توان از میان ترکیب‌های با میانگین بالا، ترکیب‌های استعاری را با دقت بالایی استخراج نمود.

در ادامه دو نمونه استعاری و نااستعاری بر اساس این روش شرح داده می‌شود.

۳-۳-۱- نمونه نااستعاری «پسر خوب»

این ترکیب یک ترکیب رایج در زبان فارسی محسوب می‌شود که فاقد معنای استعاری است.

در جدول ۱ خود ترکیب، و ۱۵ ترکیب ابتدایی برای اسم با صفت‌های نزدیک و صفت با اسم‌های نزدیک، به همراه رتبه آنان در دوگرم‌های پیکره دیده می‌شود:

جدول ۱: بررسی نمونه نااستعاری «پسر خوب»

پسر خوب: ۹۰۹۵
دختر خوب: ۱۳۷۰۰
پسری خوب: ۳۸۱۲۶
پدر خوب: ۳۵۳۱۱
عمو خوب: ۳۷۴۱۹
فرزند خوب: ۳۶۵۵۸
دختری خوب: ۳۸۰۳۲
برادر خوب: ۳۶۲۱۰
مادر خوب: ۳۱۲۵۶
پسرعمو خوب: یافت نشد.
نوه خوب: ۳۸۱۹۴
خواهر خوب: ۳۶۵۰۷
دخترک خوب: ۳۸۰۴۲
پسربچه خوب: یافت نشد.
خواهرزاده خوب: یافت نشد.
عمه خوب: ۳۸۱۱۸

[۱۴] مثلا بابت خطای ابزار، ترکیب اشتباه «بیان پاکستانی» از جمله «نماینده مردم اصفهان در مجلس با بیان اینکه ۱۳۹۸ میلیون پاکستانی شیعه هستند، گفت:» به عنوان ترکیب وصفی استخراج شده است. با محدودیت یادشده چنین مواردی حذف می‌شوند.

[۱۵] Bigram

پسر خوبی: ۳۰۱۹۷
پسر بد: ۳۶۲۶۲
پسر بهتر: ۳۷۹۴۵
پسر خوبو: ۳۸۱۸۳
پسر خب: ۳۸۱۷۳
پسر خیلی: ۳۱۱۱۰
پسر جوری: ۳۸۱۹۷
پسر اینطوری: ۳۸۱۷۱
پسر خوبن: یافت نشد.
پسر چطور: ۳۷۹۷۴
پسر هم: ۱۲۹۰۱
پسر خوش: ۳۴۴۷۸
پسر خوشحالم: یافت نشد.
پسر بهترین: ۳۷۷۰۸
پسر خداروشکر: یافت نشد.

همان‌طور که دیده می‌شود بسیاری از ترکیب‌های تولیدی، در فهرست دوگرم‌ها یافت می‌شوند. این وضعیت بر خلاف ترکیب استعاری در نمونه بعدی است.

۳-۳-۱- نمونه استعاری «تاکسی پیر»

این ترکیب یک ترکیب غیرمعمول و استعاری در زبان فارسی است. در جدول ۲ خود ترکیب، و ۱۵ ترکیب ابتدایی برای اسم با صفت‌های نزدیک و صفت با اسم‌های نزدیک، به همراه رتبه آنان در دوگرم‌های پیکره دیده می‌شود:

جدول ۲: بررسی نمونه استعاری «تاکسی پیر»

تاکسی پیر: ۳۸۲۰۴
اتوبوس پیر: ۳۸۱۵۶
راننده پیر: ۳۸۰۸۷
کرایه پیر: یافت نشد.
مترو پیر: یافت نشد.
سوار پیر: یافت نشد.
ماشین پیر: یافت نشد.
خیابون پیر: یافت نشد.
اتومبیل پیر: یافت نشد.
ماشینو پیر: یافت نشد.
تاکسیران پیر: یافت نشد.
دربست پیر: یافت نشد.
اتوبوسرانی پیر: یافت نشد.
ترمینال پیر: یافت نشد.
کامیون پیر: یافت نشد.
آسانسور پیر: یافت نشد.
تاکسی فرتوت: یافت نشد.
تاکسی سالخورده: یافت نشد.
تاکسی خرفت: یافت نشد.
تاکسی پیری: یافت نشد.
تاکسی پیرمرد: ۳۸۱۵۳
تاکسی مرد: ۳۸۱۹۶
تاکسی ترسا: یافت نشد.
تاکسی عجوزه: یافت نشد.
تاکسی چرکین: یافت نشد.
تاکسی تکیده: یافت نشد.
تاکسی جوان: ۳۸۱۹۶
تاکسی پیران: یافت نشد.
تاکسی فرسوده: ۳۸۰۸۶
تاکسی پیرزن: یافت نشد.
تاکسی میانسال: یافت نشد.

در این نمونه مشاهده می‌شود که اکثریت بالایی از ترکیب‌های تولیدی اساسا در فهرست دوگرم‌های پیکره دیده نشده است و این بدان معنا است که در کل پیکره ۵میلیارد واژه‌ای hmBlogfa، چنین ترکیبی نیامده است. این وضعیت می‌تواند حاکی از نامعمول بودن چنین ترکیبی در زبان رایج فارسی قلمداد شود و باعث شود که احتمال استعاری بودن ترکیب، بالا باشد.

۴- نتایج

الگوریتم معرفی‌شده روی تمام ترکیب‌های وصفی که در بخش ۳-۲ استخراج شده اعمال گردید. در جدول ۳ فهرستی از انواع نامزدهای استعاری بودن که طبق الگوریتم بالاترین امتیازات را گرفته‌اند مشاهده می‌شود. برای درک بهتر معنی هر ترکیب وصفی، جمله‌ای که ترکیب در آن آمده بوده است نیز به عنوان شاهد آورده شده است:

جدول ۳: اجرای الگوریتم روی ترکیب‌های وصفی استخراج‌شده

۱- وابستگی مسحور
بر این اساس خیلی راحت می توان به ماهواره و فرایند استفاده بی رویه و وابستگی مسحور کننده اش هم به عنوان یک اعتیاد مدرن نگریست.
۲- وات پنوماتیک
این پلت فرم از یک بالابر با یک جفت موتور ۱۳۹۸ وات پنوماتیک طراحی شده است.
۳- والایی شیری
میدان شهرداری، انتهای کوچه شهید والایی شیری، دانشگاه پیام نور مرکز بهارتویسرکانتویسرکان:
۴- وحشیگری مسری
با اینهمه، تمامیت ارضی روسیه و مصالح ژئوپولیتیک مسکو توجیهی است بر عملیات ضد تروریستی با وحشیگری مسری که چچن نمونه بارز آن است.
۵- وردپرس منطقی

اگر یک سایت ساده می خواهید و قصد وقت گذاشتن زیاد برای مدیریت سایت تان ندارید انتخاب وردپرس منطقی است.
۶- وفاداری متقاطع
۱۳۹۸ ظهور قرون وسطاگرایی جدید به شکل واحدهای کوچک تر با وفاداری های متقاطع
۷- ولایتی شاداب
ولایتی شاداب و سرسبز در شمال کابل و با تاکستانهای معروف.
۸- ویت گنگ
مردی که در این عکس کشته می شود، یکی از چریک های ویت گنگ است که با نیروهای آمریکایی و ویتنامی های متحد آمریکا می جنگید.
۹- ویکی پرنشاط
مال یکی حزن آور ویکی پرنشاط ویکیزندگی تاریکی داره.
۱۰- هرتز مهلک
تجربه ثابت کرده که فرکانس ۱۳۹۸ تا ۱۳۹۸ هرتز مهلک ترین فرکانس می باشد.
۱۱- هرزگی کیلومتری
و جواد برای پیشگیری از هرزگی چند کیلومتری را رکاب می زند آن هم تک پا.
۱۲- هموگلوبین توانا
رایج ترین اقدام درمانی برای تالاسمی ماژور، که البته اقدامی کنترلی است، تزریق خون سالم و دارای هموگلوبین تواناست.
۱۳- هیجان مظنون
هیچ وقت نباید از روی هیجان مظنون رو انتخاب کنی.
۱۴- هیکل فخیم
به هیکل فخیم نگاه کردم.
۱۵- یادته عبری
یادته عبری رو با الف نوشتی؟
۱۶- یقینی ملتهب
چه شاعرانه است وقتی شمع به پایان خود می رسد بی حضور گل و پروانه و شاعرانه ترچشمی که می گرید بر یقینی ملتهب از حریق ریزه باورها
۱۷- یکان معنادار
به بیان دیگر در این زبانتک واژها یکان های معنادار به همدیگر متصل می شوند تا یکواژهساخته شود.
۱۸- یوزر حدس
یک یوزر حدس بزنیدکه من طریقه حدس یوزر را در بالا گفته ام
۱۹- واد طبی
اگر پوست خود را با واد طبی سفت کنید به پیشگیری از آسیب ها کمک خواهید کرد.

همان‌طور که مشهود است برخی موارد کاملا درست و برخی نیز کاملا اشتباه هستند. مبدا اشتباهات را می‌توان در گروه‌های زیر دسته‌بندی نمود:

- موارد ناشی از اشتباهات تایپی، املایی و نگارشی نویسنده. با توجه به اینکه پست‌های وبلاگ‌ها توسط انواع نویسندگان با سطح ادبی و زمینه‌های متفاوت تولید شده‌اند، اینگونه اشتباهات عادی محسوب می‌شوند (مانند مورد ۹ در جدول ۳).
- موارد ناشی از اشتباهات خود ابزار تجزیه‌گر وابستگی، که می‌تواند ناشی از قطعه‌بندی اشتباه جملات، برچسب‌گذاری غلط جزء کلام[۱۶] و نیز کمبود داده آموزشی برای ابزار باشد (مانند موارد ۵ و ۱۱ در جدول ۳).
- اشتباهات ناشی از عامیانه یا محاوره‌ای بودن جملات، که طبعا روی نتیجه تجزیه‌گر وابستگی موثر است (مانند مورد ۱۵ در جدول ۳).
- موارد ناشی از خاص بودن سبک نوشته، مثل اشعار یا نوشته‌های کهن یا تخصصی است (مانند مورد ۱۹ در جدول ۳).
- اشتباهات ناشی از واژه‌های خاص، نام‌های موجودیت‌های نام‌دار[۱۷] که در فارسیِ رایج و پیکره‌های محدود دیده نمی‌شوند؛ یا اگر باشند دارای ابهام هستند، و همگی این موارد روی دقت ابزار تاثیر می‌گذارند (مانند موارد ۳ و ۸ در جدول ۳).

البته باید دقت داشت که چون روش پیشنهادی این مقاله در پی کشف ناهنجاری در ترکیب‌ها می‌باشد، طبیعتا موارد یادشده در بالا نیز، که در واقع نوعی ناهنجاری می‌باشند در صدر فهرست خروجی تجمع یافته‌اند. به زبان دیگر، این که چنین اشتباهاتی در میان خروجی‌ها دیده شود، بیش از آنکه نقص روش پیشنهادی تلقی گردد، نشان‌گر نقص در صحت داده و ابزارهای پایه می‌باشد که بدین شکل بر خروجی روش اثر گذارده است.

در این راستا و برای کاستن از مواردی این چنین، دو تصمیم در نظر گرفته شد. یکی استفاده از برخی قواعد ساده که می‌توان با کمک آنان به مقابله با برخی الگوهای اشتباه‌زا رفت. مثلا تعداد زیادی از ترکیب‌های تولیدی اشتباه، ترکیب‌هایی است که بابت اشتباه ابزار، موصوف، کلماتی شبیه «ها» و «های» هستند می‌باشد.

در تصمیم دوم بنا بر این گذاشته شد که ترکیب‌های وصفی‌ای مد نظر قرار گیرند که رایج‌تر و پرتکرارتر محسوب می‌شوند. برای این کار، رتبه موصوف آن‌ها از نظر رخداد در پیکره[۱۸]، مبنا قرار گرفت. بدین ترتیب برخی از مواردی که ناشی از اشتباهات نادر و فاحش نویسنده یا ابزار می‌باشند

[۱۶] POS Tagging

[۱۷] مثل مواردی چون «صابر صیقلی»، «اسکندر حقیقی»، «بازرگانی درخشان» و ...

[۱۸] قرار گرفتن در پنج هزار واژه پررخداد پیکره

حذف گردیدند[۱۹]. بدین ترتیب مواردی مثل «یوزر حدس»، «یادته عبری»، «وات پنوماتیک » و ... (از جدول ۳) حذف گردیدند.

با اعمال موارد بالا، خروجی اجرایی به دست آمده روی همه ترکیب‌های وصفی معادل ۵۴۴۶۲۶ ترکیب می‌باشد. البته از این میزان، ۱۹۱۹۱۸ترکیب، ترکیب‌هایی بودند که دست‌کم یکی از اجزای صفت یا موصوف، در مدل جاسازی واژه‌ها یافت نشده است و به همین دلیل حذف گردیده‌اند. این وضعیت ناشی از آن بوده است که مدل زبانی بر اساس پیکره ریشه‌یابی‌شده آموزش یافته است؛ در حالی که جملاتی که تحویل تجزیه‌گر وابستگی شده‌اند تا ترکیب‌های وصفی از آنان استخراج گردد، ریشه‌یابی نشده‌اند. برای همین طبیعی است اجزای ترکیب‌هایی مثل «کارهای ترسناک» یا «ارتباطات فضایی» اساسا در مدل یافت نشوند[۲۰].

این تصمیم دارای دو وجه بوده است. در وجه اول و بر اساس الگوریتم پیشنهادی برای استخراج استعاره‌ها، برابر بررسی‌های انجام‌شده بهتر آن بوده است که واژه‌ها به صورت ریشه‌یابی‌شده در مدل آموزش ببینند. به عنوان نمونه بهتر است واژه‌های «برادران، برادرها، برادرهای، برادرم و ...» همگی «برادر» در نظر گرفته شوند و در یک فاصله با «پسر» قرار گیرند[۲۱]. در وجه دوم و با توجه به خطاهای قابل‌توجه در خروجی تجزیه‌گر وابستگی، احتمالا بهتر است که از دست‌کاری بیشتر جمله‌ها و ریشه‌یابی اجزای آن خودداری شود. چراکه جمله‌هایی که ریشه‌یابی می‌شوند از حالت عادی بیان فارسی‌زبانان خارج می‌شود و این می‌تواند اثر سوء بر تجزیه‌گر گذارد.

بدین ترتیب تعداد نهایی ترکیب‌هایی که الگوریتم استخراج استعاره روی آن اجرا گردیده است ۳۵۲۷۰۸ می‌باشد. نمره میانگین در نااستعاری‌ترین حالت ۲۱۵۸۰/۳۲ و در استعاری‌ترین حالت ۳۸۲۰۸ می‌باشد.

از آنجا که هیچ داده فارسی برای بررسی استعاره یافت نشده است برای ارزیابی نتایج و حذف موارد اشتباه، از ارزیابی انسانی استفاده شده است. جزئیات بیشتر در بخش ۵ آمده است.

۵- ارزیابی و خروجی نهایی

همان‌طور که ذکر شد خروجی الگوریتم فهرست بلندی است از ترکیب‌های وصفی، که امتیازی برای میزان استعاری بودن هر ترکیب به آنان منتسب شده است. برای بررسی این خروجی، و با توجه به نبود داده برچسب‌خورده فارسی در زمینه استعاره، چاره‌ای جز بررسی انسانی کار باقی نمی‌ماند.

طبیعتا بررسی دستی تمام بیش از ۳۵۰هزار ترکیب وصفی، امر ساده‌ای نیست. برای همین و با توجه به محدودیت زمان و امکانات، تنها ۲۱۲۵ مورد از خروجی، و به ترتیب از استعاری‌ترین ترکیب‌ها مورد بررسی قرار گرفتند.

فرآیند بررسی به این صورت تنظیم شد که یک ترکیب وصفی، در جمله بررسی می‌شود. برای هر ترکیب حالت‌های زیر در نظر گرفته شده است:

- استعاری: ترکیب استعاری است. مثل «قفس لاابالی» در جمله «من به فکر های پریده از قفس لاابالی مغزم نمی رسم».
- نااستعاری: ترکیب نااستعاری است. مثل «موبایل فولادی» در جمله «گوشی موبایل فولادی برای افراد ثروتمند».
- مشکوک: ترکیب مشکوک به استعاری بودن است. مثل «شغل بزمی» در جمله «یک شغل رزمی و یک شغل بزمی».
- نامعلوم: جمله برای کاربر بررسی‌کننده نامفهوم بوده است و معنای آن را نفهمیده است. مثل «بال مزدوری» در جمله «آدم های بدجنس باید از کارها اخراج شوند و بال مزدوری کتم کرنگه!».
- اشتباه: ترکیب وصفی اساسا اشتباه تشخیص داده شده است و اصولا ترکیب وصفی صحیحی نیست. این مورد معمولا ناشی از خطای تجزیه‌گر و یا اشتباهات نگارشی و املایی است. مثل «آمار موزاییک» در جمله «انگار قسمت نبود ، آمار موزاییک های خیابان مورد نظر در دست با کفایت ما باشد».

نتایج بررسی و ارزیابی در جدول ۴ آمده است:

جدول ۴: نتایج برچسب‌گذاری خروجی

نوع	تعداد
استعاری	۳۱۲
نااستعاری	۵۵
مشکوک	۳۰۰
نامعلوم	۲۲۵
اشتباه	۱۲۳۳
مجموع	۲۱۲۵

برای محاسبه دقت چاره‌ای نیست که سهم جملات با برچسب «اشتباه» و «نامعلوم» حذف شوند. چراکه خطاهای ناشی از این جملات خطاهای مربوط به روش پیشنهادی مقاله نیست. بر این اساس، در جدول ۵ دقت روش بر مبناهای مختلف آمده است:

جدول ۵: دقت نتایج

مبنا	دقت
با در نظر گرفتن مشکوک‌ها به عنوان نااستعاری	٪۴۷
با در نظر گرفتن مشکوک‌ها به عنوان استعاری	٪۹۲
با حذف مشکوک‌ها	٪۸۵

خروجی نهایی در قالب یک فایل اکسل در نشانی:
http://fs.nlp.sbu.ac.ir/members/motahari/metr/papers/metaphoric-noun-adjective-detection

[۱۹] البته طبعا در این بین ممکن است برخی موارد صحیح نیز از دست رفته باشند. ولی برای هرس کردن و کاهش نتایج غلط، در مجموع اقدامی سودمند می‌باشد.

[۲۰] ریشه‌یاب استپ‌وان این ترکیب‌ها را تبدیل به «کار ترسناک» و «ارتباط فضایی» می‌کند.

[۲۱] برای بررسی ترکیبی مثل «پسر خوب» که در ۳-۳-۱ آمد.

قابل دستیابی است. در این فایل اکسل، و در هر ردیف، خود ترکیب وصفی، برچسب آن («استعاری»، «نااستعاری» و ...)، امتیاز آن و جمله شاهد برای ترکیب آمده است.

این نتایج با توجه به پیچیدگی شناخت استعاره‌ها، حتی توسط عامل‌های انسانی، به نظر بسیار مطلوب می‌نماید. به ویژه آنکه می‌توان گفت اساسا ترکیب‌های نااستعاری وصفی در زبان، به طور قابل‌ملاحظه‌ای بیشتر از ترکیب‌های وصفی استعاری می‌باشند و این میزان از دقت برای استخراج ترکیب‌های استعاری دستاوردی ارزشمند است.

۶- جمع‌بندی

این مقاله در پی معرفی یک روش برای استخراج ترکیب‌های وصفی استعاری در زبان فارسی با کمک گرفتن از پیکره‌های بزرگ زبانی و مدل جاسازی واژه‌ها بود. با توجه به کمبود داده استعاری در زبان فارسی و نیز پیچیدگی مفهوم استعاره در زبان، تهیه هر نوع داده برچسب‌خورده استعاری غنیمت است و این پژوهش در این راستا یک گام رو به جلو محسوب می‌شود.

همان‌طور که دیده شد، یکی از مشکلات عمده در استفاده از روش پیشنهادی این مقاله، کیفیت نامطلوب ابزارهای پایه پردازش زبان طبیعی در تجزیه نحوی و وابستگی، در زبان فارسی است. این کیفیت نامطلوب دست‌کم در خروجی روش این مقاله به شدت خود را نشان می‌دهد و ضرورت بهبود در ابزارهای پایه برای زبان فارسی را یادآور می‌شود.

علاوه بر این، بخش بزرگی از مفاهیم استعاری در زبان، در ترکیب‌هایی به غیر از ترکیب‌های وصفی خود را نشان می‌دهند و به عنوان کارهای پیش رو، تهیه داده برچسب‌خورده استعاری به صورت عمومی، می‌تواند یک هدف مهم برای کارهای آینده باشد. البته به عنوان یک نکته دیگر، حالت‌های پیچیده‌تر استعاره لزوما به صورت تک‌واژه‌ای یا چند واژه پشت سر هم رخ نمی‌دهند، و این خود می‌تواند پژوهش در این حوزه را دشوارتر سازد.

مراجع

Birke, J., & Sarkar, A. (2006). A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language. *EACL*.

Bizzoni, Y., & Ghanimifard, M. (2018). Bigrams and BiLSTMs Two neural networks for sequential metaphor detection. *Proceedings of the Workshop on Figurative Language Processing*, 91–101.

Colston, H. L. (2015). *Using figurative language*. New York, NY: Cambridge University Press.

Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge University Press.

Dunn, J. (2013). Evaluating the premises and results of four metaphor identification systems. *International Conference on Intelligent Text Processing and Computational Linguistics*, 471–486. Springer.

Fass, D. (1991). met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, *17*(1), 49–90.

Gedigian, M., Bryant, J., Narayanan, S., & Ciric, B. (2006). Catching metaphors. *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, 41–48. Association for Computational Linguistics.

Gershman, Y. T. L. B. A., & Dyer, E. N. C. (2014). Metaphor detection with cross-lingual model transfer. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Haagsma, H., & Bjerva, J. (2016). Detecting novel metaphor using selectional preference information. *Proceedings of the Fourth Workshop on Metaphor in NLP*, 10–17.

Jia, Y., & Yu, S. (2008). Unsupervised Chinese Verb Metaphor Recognition Based on Selectional Preferences. *PACLIC*, 207–214.

Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing* (2nd ed.). Upper Saddle River, N.J: Prentice Hall.

Klebanov, B. B., Leong, C. W., & Flor, M. (2015). Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. *Proceedings of the Third Workshop on Metaphor in NLP*, 11–20.

Kövecses, Z., & Benczes, R. (2010). *Metaphor: A practical introduction* (2nd ed). New York: Oxford University Press.

Krishnakumaran, S., & Zhu, X. (2007). Hunting elusive metaphors using lexical resources. *Proceedings of the Workshop on Computational Approaches to Figurative Language*, 13–20. Association for Computational Linguistics.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Mason, Z. J. (2004). CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, *30*(1), 23–44.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Howard, N., & Frieder, O. (2013). Metaphor identification in large texts corpora. *PloS One*, *8*(4), e62343.

Pramanick, M., & Mitra, P. (2018). Unsupervised Detection of Metaphorical Adjective-Noun Pairs. *Proceedings of the Workshop on Figurative Language Processing*, 76–80.

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018). Universal Dependency Parsing from Scratch. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 160–170. Retrieved from https://nlp.stanford.edu/pubs/qi2018universal.pdf

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., … Assi, S. M. (2010). Semi automatic development of farsnet; the persian wordnet. *Proceedings of 5th Global WordNet Conference, Mumbai, India*, *29*.

Shamsfard, M., Jafari, H. S., & Ilbeygi, M. (2010). STeP-1: A Set of Fundamental Tools for Persian Text Processing. *LREC*.

Shutova, E. (2016). Design and evaluation of metaphor processing systems. *Computational Linguistics*.

Shutova, E., & Sun, L. (2013). Unsupervised Metaphor Identification Using Hierarchical Graph Factorization Clustering. *HLT-NAACL*, 978–988.

Shutova, E., Sun, L., & Korhonen, A. (2010). Metaphor identification using verb and noun clustering. *Proceedings of the 23rd International Conference on Computational Linguistics*, 1002–1010. Association for Computational Linguistics.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., & Krennmayr, T. (2010). *VU Amsterdam Metaphor Corpus*. University of Oxford Text Archive. http://ota. ahds. ac. uk/desc/2541.

Su, C., Huang, S., & Chen, Y. (2017). Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, *219*, 300–311. https://doi.org/10.1016/j.neucom.2016.09.030

Swarnkar, K., & Singh, A. K. (2018). Di-lstm contrast: A deep neural network for metaphor detection. *Proceedings of the Workshop on Figurative Language Processing*, 115–120.

Turney, P. D., Neuman, Y., Assaf, D., & Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 680–690. Association for Computational Linguistics.

Wilks, Y., Dalton, A., Allen, J., & Galescu, L. (2013). Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. *Proceedings of the First Workshop on Metaphor in NLP*, 36–44.

Wu, C., Wu, F., Chen, Y., Wu, S., Yuan, Z., & Huang, Y. (2018). Neural metaphor detecting with cnn-lstm model. *Proceedings of the Workshop on Figurative Language Processing*, 110–114.

Yooz search engine. (n.d.). Retrieved from http://yooz.ir/