تعیین مشابهت معنایی به روش بدون‌سرپرست با استفاده از قدم‌زنی تصادفی بر گراف جایگزینی زبانی

فاطمه کاوه‌یزدی[×۱]، دانشجوی دکتری؛ علی‌محمد زارع‌بیدکی[†۲]، دانشیار؛ محمدرضا پژوهان[۳]، استادیار

۱- گروه مهندسی کامپیوتر - دانشگاه یزد - یزد - ایران - fkavehy@stu.yazd.ac.ir
۲- گروه مهندسی کامپیوتر - دانشگاه یزد - یزد - ایران - alizareh@yazd.ac.ir
۳- گروه مهندسی کامپیوتر - دانشگاه یزد - یزد - ایران - pajoohan@yazd.ac.ir

چکیده: این مقاله به معرفی روشی برای تعیین مشابهت معنایی کلمات با استفاده از پیکره‌های تنک می‌پردازد. این روش با ارائه مفهوم جایگزین‌پذیری غیرمستقیم برای اولین بار و پیاده‌سازی آن از طریق گراف جایگزین‌پذیری عبارت‌ها توانسته است بر مشکل تنک بودن فضای زمینه در زبان‌های با منابع محدودتر مانند فارسی غلبه نماید. از طرف دیگر باید به این نکته اشاره نمود که برای تولید گراف جایگزینی لازم برای تعیین مشابهت معنایی می‌توان از پیکره‌های متنی به صورت مستقل از زبان بهره گرفت. نتایج ارزیابی‌ها با استفاده از دادگان آزمون مجموعه RG-65 که از دادگان متداول برای ارزیابی کیفیت تعیین مشابهت معنایی است، نشان می‌دهد که مقدار ضریب همبستگی Spearman این روش بین ۰/۰۳ تا ۰/۱۳ واحد بیش از سایر روش‌های بدون سرپرست موفق است.

واژه‌های کلیدی: مشابهت معنایی، جایگزینی زبانی، گراف جایگزینی، قدم‌زنی تصادفی، پیکره، ویکی‌پدیا.

Unsupervised Semantic Similarity Estimation using Random Walk on Lexical Substitution Graph

F. Kaveh-Yazdy[1], PhD Student; A. M. Zareh-Bidoki[2], Associate Professor; M. R. Pajoohan[3], Assistant Professor

1- Department of Computer Engineering, Yazd University, Yazd, Iran, fkavehy@stu.yazd.ac.ir
Senior Researcher at Parsijoo Persian Search Engine, fkavehy@parsijoo.ir
2- Department of Computer Engineering, Yazd University, Yazd, Iran, alizareh@yazd.ac.ir
Head of Parsijoo Persian Search Engine, alizareh@parsijoo.ir
3- Department of Computer Engineering, Yazd University, Yazd, Iran, pajoohan@yazd.ac.ir

Abstract: This paper introduces the indirect substitutability relation for the first time to provide a practical solution for estimating semantic similarity. Proposed method is an unsupervised semantic similarity estimation method, which is benefitted from taking into account the indirect substitutability relation. This method recognizes the substitutability between two terms by considering a third term, which has similar lexical context with each of them separately. To model this relation, we generate a graph using substitutable pairs of terms. The strength of the relation between each pair of terms is approximated by propagating semantic score through the substitutability graph. This method is language independent and uses only textual corpora to generate the substitution graph. Furthermore, it supports semantic similarity estimation in languages suffering from lack of dense corpora. Results of our experiments using RG-65 Persian dataset show that the proposed method outperforms the baseline algorithms. The proposed method improves the estimation from 0.03 Spearman's correlation up to 0.13 in comparison with the baseline algorithms.

Keywords: Semantic similarity, lexical substitution, substitution graph, random walk, corpus, wikipedia.

تاریخ ارسال مقاله: ۱۳۹۵/۰۶/۱۵
تاریخ اصلاح مقاله: ۱۳۹۵/۰۹/۰۲
تاریخ پذیرش مقاله: ۱۳۹۵/۱۰/۲۲
نام نویسنده مسئول: محمدرضا پژوهان
نشانی نویسنده مسئول: ایران - یزد - بلوار پژوهش - دانشگاه یزد - گروه مهندسی کامپیوتر.

× محقق ارشد در موتور جستجوی پارسی‌جو
† مدیر پروژه موتور جستجوی پارسی‌جو

۱- مقدمه

روابط معنایی بین کلمات از دیرباز در پردازش‌های زبانی مورد توجه قرار گرفته‌اند و از شاخص‌ترین این روابط می‌توان به رابطه ترادف اشاره نمود که بین هر دو وجه معنایی از دو کلمه قابل تعریف است. قدیمی‌ترین تعریف رسمی از دو کلمه مترادف توسط لایبنیز ریاضیدان شهیر آلمانی ارائه شده است [۱]. لایبنیز دو کلمه را مترادف می‌داند، هرگاه با جایگزینی یکی از آن‌ها با دیگری، حقیقت بیان شده در جمله هرگز تغییر نکند. بر اساس این تعریف تعداد بسیار کمی از کلمات می‌توانند به‌عنوان مترادف قلمداد شوند. این تعریف بر دو اصل اساسی تکیه می‌کند: (الف) جایگزین‌پذیری۱ و (ب) زمینه زبانی۲ مشترک. در واقع لایبنیز، ترادف را به صورت نسبی و با توجه به توانایی جایگزین کردن دو کلمه در یک زمینه زبانی مشترک تعریف می‌کند. بر همین اساس یک تعریف ضعیف‌تر از ترادف را می‌توان به صورت زیر ارائه نمود[۱]:

«دو کلمه در یک زمینه زبانی مانند C مترادف هستند، هرگاه جایگزینی آنها در زمینه C تغییری در حقیقت بیان شده ایجاد نکند.»

با این اغماض، فرایند تعیین مترادف تسهیل شده است و یافتن حتی یک زمینه مشترک می‌تواند به تعیین رابطه ترادف بین دو کلمه بیانجامد. در راستای رسمیت بخشیدن به تعریف رابطه ترادف، زبان‌شناسان رابطه کلی‌تری به نام رابطه "مشابهت معنایی"[۳] را تعریف نموده‌اند. این رابطه در ابتدا در سطح کلمات تعریف شد و براساس آن، برای هر جفت کلمه براساس میزان جایگزین‌پذیری آن‌ها می‌توان مشابهت معنایی با مقدار شدّت متفاوت درنظر گرفت. میزان شدّت این رابطه برای کلمات مترادف دارای بالاترین مقدار و برای کلمات کاملاً بی‌ربط دارای مقدار صفر است. استفاده از این رابطه توانست گام مؤثّری در تولید خودکار هستان‌شناسی‌ها و واژه‌نامه‌ها بردارد [۲] و با مرور زمان نیاز به استفاده از آن در زمینه‌های متعدّد احساس شد. این رابطه می‌توانست بین کلمات، عبارت‌ها، جملات، پاراگراف‌ها و یا حتی اسناد تعریف شود و به همین دلیل استفاده از آن در سیستم‌های بازیابی اسناد مانند [۵-۳]، سیستم‌های ترجمه ماشینی [۶] و سیستم‌های پرسش و پاسخ [۷] مورد توجه قرار گرفت.

این مقاله روشی را معرفی می‌کند که هیچ وابستگی به زبان یا کاربر انسانی ندارد. برای نیل به این هدف، برای اولین بار رابطه جایگزین‌پذیری غیرمستقیم تعریف شد که نتایج آن به ارائه چهارچوبی برای تعیین شدت جایگزین‌پذیری با استفاده از عبارت واسطه منتهی شد. انعطاف بیشتر برای استخراج جایگزینی در پیکره‌های با اندازه متوسط و غنای کم و قابلیت پیاده‌سازی بر روی بسترهای محاسباتی توزیع شده از ویژگی‌های دیگر چهارچوب پیشنهادی است.

در ادامه این مقاله، روش‌های محاسبه مشابهت معنایی به صورت اجمالی در بخش دوم برشمرده خواهند شد. در بخش سوم، تعریفی رسمی از جایگزین‌پذیری به عنوان راهکاری برای تعیین مشابهت معنایی ارائه می‌گردد و در ادامه این بخش فرایند محاسبه مشابهت معنایی برای یک جفت کلمه تشریح می‌گردد. جزئیات مربوط به تولید دادگان و پیاده‌سازی روش‌های انتخاب شده برای مقایسه در بخش چهارم بیان خواهد شد و در نهایت جمع‌بندی اجمالی از مطالب و نتیجه‌گیری در بخش پنجم منعکس خواهد شد.

۲- مروری بر روش‌های تعیین مشابهت معنایی

سیستم‌های بازیابی اطلاعات در بخش‌هایی مانند گسترش پرس‌وجو (مانند [۸]) و سیستم‌های ترجمه ماشینی هر کدام به نوعی با مقوله تعیین مشابهت معنایی سروکار داشته‌اند. این سیستم‌ها دو رویکرد اصلی در تعیین مشابهت معنایی را مورد استفاده قرار داده‌اند:

۱- تعیین مشابهت معنایی با استفاده از منابع زبان‌شناختی

۲- تعیین مشابهت معنایی با استفاده از پیکره

عمده تحقیقاتی که در دسته (۱) قرار می‌گیرند، از واژه‌نامه‌ها و یا هستان‌شناسی‌ها، مانند WordNet یا نمونه‌های مشابه با آن در زبان‌های دیگر، استفاده می‌کنند. غالب این تحقیقات روش‌های متنوعی را برای محاسبه فاصله معنایی کلمات در دادگان زبانی خود مورد استفاده قرار داده‌اند؛ که از آن جمله می‌توان به تحقیقاتی مانند [۱۱-۹] اشاره کرد. در تعداد قابل توجهی از این تحقیقات، WordNet به مانند یک گرافِ دربرگیرنده روابط معنایی مورد استفاده قرار گرفته است و فاصله هر یک از معانی کلمات با دنبال کردن مسیر روابط موجود در این گراف معنایی تعیین شده است [۱۴-۱۲]. باید خاطرنشان نمود که گرچه دقت روش‌های مبتنی بر هستان‌شناسی قابل توجه است، اما این روش‌ها در برخورد با کلمات جدیدی که در هستان‌شناسی موجود نباشند به مشکل برخواهند خورد.

به دنبال مشکلات دسته اوّل، محققین تصمیم گرفتند از پیکره‌های زبانی برای تعیین میزان شباهت کلمات استفاده نمایند. این روش‌ها با اتکا بر این فرض که می‌توان در حجم انبوهی از متن، انواع جایگزینی‌های ممکن از دو کلمه را در زمینه‌های یکسان دید، کار خود را شروع کردند. رویکردهای مورد استفاده در روش‌های مبتنی بر استفاده از پیکره را می‌توان به دو دسته تقسیم نمود که عبارتند از:

الف) روش‌های مبتنی بر معیارهای ساده

ب) روش‌های مبتنی بر فضای برداری

روش‌های دسته (الف) سعی نموده‌اند با تعریف معیارهایی، عمدتاً مبتنی بر هم‌وقوعی کلمات، میزان نزدیکی معنایی آنها با یکدیگر را بسنجند. از نظر این روش‌ها، مجموعه کلماتی که در پنجره‌هایی با طول محدود ظاهر شده باشند از نظر معنایی با هم مرتبط هستند. این تعریف لزوماً تاکیدی بر مترادف بودن و قابلیت جایگزینی کلمات با یکدیگر ندارد و تنها وابستگی معنایی را مورد توجه قرار می‌دهد. یکی از متداول‌ترین معیارهای بکار رفته در این دسته از روش‌ها، معیار میزان اطلاع توأم نقطه‌ای[۴] است [۱۵]. البته انواع دیگری از این معیار نظیر اطلاع توأم منطقه‌ای مثبت نیز در این زمینه مورد توجه قرار گرفته‌اند [۱۶]. جدیدترین نمونه از این دسته از معیارها، معیار فاصله

نرمال گوگل است که توسط Cilibrasi و Vitanyi [۱۷] پیشنهاد شده است و از دادگان چندجزیی گوگل[۵] استفاده می‌کند.

روش‌های معرفی شده در گروه الف، می‌توانستند میزان ارتباط معنایی بین کلماتی که در یک متن در کنار هم و یا با فاصله مناسبی (کوچک‌تر از طول پنجره انتخاب شده) واقع شده بودند را محاسبه نمایند، امّا کماکان در تعیین کلمات مترادف ناکارآمد بودند. زیرا در غالب موارد دو کلمه هم‌معنا و قابل جایگزینی در کنار هم در یک پنجره کوتاه ظاهر نمی‌شوند. بنابراین روش‌های دسته اوّل، گاه فاصله معنایی زیادی برای کلمات مترادف تخمین می‌زدند.

برای حل این مشکل محققین مجدداً تلاش‌های خود را بر روی روش‌های تخمین شباهت معنایی مبتنی بر جایگزین‌پذیری در زمینه یکسان معطوف نمودند. نسل اوّل این روش‌ها از اواخر دهه هشتاد میلادی مورد توجه قرار گرفتند. این روش‌ها ماتریس وقوع مجموعه کلمات در متن‌ها را تشکیل می‌دهند و سپس سعی می‌کنند کلمات را براساس هم‌وقوعی در اسناد مشابه دسته‌بندی نمایند و از طرفی اسناد را نیز با توجه به مشابهت کلمات آنها دسته‌بندی کنند و این دسته‌بندی را با استفاده از تجزیه مقادیر تکین[۶] انجام می‌دهند. با استفاده از این تبدیل و محاسبه بردارهای مقادیر ویژه، ابعاد فضای مسئله در وهله اوّل کاهش می‌یابد و در وهله دوم، روابط بین کلمات و مفاهیم مندرج در متن‌ها استخراج می‌گردد. از نمونه‌های این خانواده از روش‌ها می‌توان به تحلیل معنایی نهان[۷] [۱۹، ۱۸]، تخصیص نهان دیریکله [۲۰] و تحلیل مستقل اجزا [۲۱] اشاره نمود. مجموعه این روش‌ها نیازمند ماتریس‌های بزرگ هم‌وقوعی هستند که باید در حافظه نگهداری شوند و علاوه بر آن محاسبات مربوط به استخراج بردارهای ویژه نیز بسیار پیچیده هستند.

نسل دوم از این روش‌ها نیز نمایش‌های برداری از کلمات ارائه می‌نمودند که عمدتاً به صورت افزایشی توسط یک شبکه عصبی تولید می‌شدند. این روش‌ها دسته‌ای از مدل‌های تعیین مشابهت معنایی را ایجاد کردند که به مدل‌های معنایی توزیع شده[۸] معروف شده‌اند و با استفاده از پیکره‌های زبانی و با اتکا به فرضیه Harris [۲۲] آموزش می‌بینند. براساس فرضیه Harris میزان شباهت بین دو کلمه را می‌توان براساس تابعی از درجه همپوشانی بین زمینه زبانی دو کلمه مدل نمود. بنابراین این روش‌ها تلاش می‌کنند از زمینه‌های مشترک بین کلمات که در تمام پیکره توزیع شده و در مکان‌های مختلف واقع شده‌اند برای آموزش استفاده کنند.

معروف‌ترین نمونه از این دسته توسط Mikolov و همکاران [۲۳] در گوگل پیشنهاد شد. این بازنمایی با استفاده از اطلاعات هم‌وقوعی کلمات و به صورت افزایشی توسط یک شبکه عصبی ماشین بولتزمان محدود تولید می‌شد و می‌توانست میزان ارتباط معنایی نسبی کلمات را تعیین کند. این مدل Word2Vec نامیده شد و Mikolov توانست با همکاری Le [۲۴] در مدت کوتاه نسخه تعمیم یافته این مدل را نیز ارائه نماید که می‌توانست نمایش برداری جملات، پاراگراف‌ها و حتی اسناد را محاسبه کند. بدیهی است که فرایند آموزش این شبکه عصبی با استفاده از پیکره‌های بزرگ متنی می‌تواند پرهزینه و زمان‌بر باشد. در فاصله کوتاهی از معرفی Word2Vec، Pennington و همکاران مدل GLOVE [۲۵] را ارائه دادند. این مدل با استفاده از یک مدل رگرسیون لگاریتمی بازنمایی برداری کلمات را استخراج می‌نمود. هدف از ارائه این روش غلبه بر مشکلات فاکتورگیری سراسری در روش‌هایی مانند LSA، توأم با حل مشکلات زمینه با پنجره محدود در Word2Vec بود [۲۶].

مدل‌های GLOVE و Word2Vec هر دو از دادگان بدون برچسب برای آموزش بهره می‌گرفتند و نتایج آنها دربرگیرنده خطاهایی در تشخیص ارتباط کلمات با زمینه تنک بود. زیرا غنای ماتریس هم‌وقوعی به نحو چشم‌گیری بر دقت این روش‌ها اثر دارد. برای حل این مشکل، پیشنهاددهندگان امکان استفاده از پیکره‌های با ابعاد ترابایت را مطرح نمودند که می‌توانست غنای لازم را فراهم آورد ولی همزمان فرایند آموزش شبکه عصبی را با مشکلات مدیریت حافظه دست به گریبان می‌نمود. برای رفع این مشکل محققین ترجیح دادند، انواع دیگری از ویژگی‌های زبانی و غیرزبانی را از پیکره‌ها و دادگان استخراج نموده و با بازنمایی برداری ترکیب نمایند. به عنوان مثال، Szarvas و همکاران [۲۷] با استفاده از ویژگی‌های غیر کلمه‌ای یک شبکه عصبی را به صورت با سرپرست آموزش دادند. این روش می‌تواند بر مشکل نویز ناشی از لحاظ نمودن تمام زمینه‌های موجود در پیکره غلبه کند. در تحقیق دیگری، Melamud و همکاران [۲۸] به‌جای استفاده از زمینه مشترک برای تولید بردارهای کلمات، به ازای هر زمینه مشترک، یک بردار برای هر کلمه تولید می‌کنند که شامل کلمات جایگزین و امتیاز مربوط به میزان احتمال جایگزینی آنها با کلمه هدف در این زمینه زبانی خاص است. Roller و Erk [۲۹] با ارائه تعریف ساده‌تری از آن‌چه در [۲۸] معرفی شده است، سرعت آموزش و کارایی این مدل را افزایش دادند. به صورت کلی می‌توان گفت، مجموعه این تلاش‌ها معطوف ارائه روش‌هایی است که تا حدّ ممکن از دادگان زبان‌شناختی مستقل باشند و در عین‌حال از اطلاعات کاربردی کلمات در زمینه‌های مختلف بهره بگیرند.

۳- روش پیشنهادی برای تعیین مشابهت معنایی

در این تحقیق با تأکید بر نتایج به‌دست‌آمده از تحقیقات مبتنی بر تعیین مشابهت معنایی با استفاده از پیکره، تلاش شده است تا یک روش مبتنی بر پیکره دادگان متنی با استفاده از یک مدل بدون‌سرپرست به دست آید. این روش با اتکا بر این اصل که هر دو کلمه قابل جایگزینی در یک زمینه زبانی مشابه دارای مشابهت معنایی هستند، تلاش می‌کند مجموعه کلمات و عبارات مندرج در زمینه‌های مشابه را استخراج نموده و سپس براساس میزان جایگزین‌پذیری آن‌ها با همدیگر، گراف رابطه جایگزینی را بسازد. گراف جایگزینی امکان انتقال اثر زمینه را به صورت زنجیروار فراهم می‌کند. در این بخش،

ابتدا تعاریف لازم برای ساخت گراف جایگزینی معرفی شده و در ادامه مراحل اجرای الگوریتم بیان خواهد شد.

۳-۱- تعریف جایگزین‌پذیری

در این بخش مفهوم جایگزین پذیری با اتکا به تعریف زمینه مشترک تعریف می‌شود. بر همین اساس برای هر عبارت مانند t می‌توان زمینه‌ای مانند C در پنجره‌ای به اندازه $2k+1$ تعریف کرد. یک عبارت می‌تواند شامل یک کلمه و یا بیشتر باشد، یعنی

$$t = w_1\ w_2\ \dots\ w_n \qquad n \geq 1 \tag{۱}$$

که در آن هر W_i یک کلمه است که با "فاصله" از کلمه مجاور خود جدا می‌شود. در این چهارچوب، زمینه C، برای عبارت x که در بردارنده t_i است در قالب کلمات کناری تعریف می‌شود (رابطه (۲)).

$$x = \langle t_{i-k} t_{i-k+1} \dots t_{i-1}\ t_i\ t_{i+1} \dots t_{i+k} \rangle \tag{۲}$$

براساس رابطه (۲) زمینه یک عبارت را می‌توان به صورت رابطه (۳) تعریف نمود.

$$C = \langle t_{i-k} t_{i-k+1} \dots t_{i-1}\ , t_{i+1} \dots t_{i+k} \rangle \tag{۳}$$

بر این اساس حداکثر k عبارت در طرفین عبارت موردنظر به عنوان زمینه مورد استفاده قرار می‌گیرند. باید خاطرنشان نمود که هر عبارت موجود در زمینه نیز می‌تواند شامل یک یا چند کلمه باشد. برای نمایش توأم هر زمینه و عبارت موردنظر از رابطه (۴) بهره گرفته می‌شود.

$$\langle C, t_i \rangle = \langle\langle t_{i-k} t_{i-k+1} \dots t_{i-1}\ , t_{i+1} \dots t_{i+k} \rangle, t_i \rangle \tag{۴}$$

براساس رابطه (۴)، دو عبارت t_i و t_j قابل جایگزینی هستند، اگر و فقط اگر دارای حداقل یک زمینه مشترک مانند C باشند. به بیان بهتر، رابطه جایگزینی دو عبارت، $S(t_i, t_j)$، به صورت مندرج در رابطه (۵) تعریف می‌شود.

$$iff\ \langle C, t_i \rangle \,\&\, \langle C, t_j \rangle \rightarrow S(t_i, t_j) \tag{۵}$$

با توجه به رابطه (۵) جایگزین شدن دو عبارت در یک زمینه متنی مشترک به معنای قابلیت جایگزینی آنها است. این تعریف در پیکره‌های بسیار بزرگ که احتمال بروز کلمات با مشابهت معنایی در موقعیت‌های یکسان بیشتر است، می‌تواند تعریف مناسبی باشد امّا همین تعریف در پیکره‌های تنک و یا کوچک کارا نخواهد بود. برای غلبه بر این مشکل، در این مقاله و برای اولین بار، مفهوم جایگزین‌پذیری غیرمستقیم برای محاسبه میزان مشابهت معنایی تعریف شده است. بنابراین جایگزین‌پذیری به دو رابطه تقسیم می‌شود که از این به بعد، روابط جایگزین‌پذیری مستقیم[۹] و جایگزین‌پذیری غیرمستقیم[۱۰] نامیده می‌شوند. رابطه جایگزین‌پذیری مستقیم در رابطه (۶) مشابه رابطه جایگزین پذیری (رابطه (۵)) تعریف می‌شود.

$$iff\ \langle C, t_i \rangle \,\&\, \langle C, t_j \rangle \rightarrow DS(t_i, t_j) \tag{۶}$$

رابطه جایگزین‌پذیری غیرمستقیم بین دو عبارت غیریکسان تعریف می‌شود که هر کدام با دو زمینه متفاوت جایگزین مستقیم یک عبارت ثالثِ واحد شده باشند. در این چهارچوب، روابط زیرمجموعه جایگزین‌پذیری به صورت رابطه (۷) تعریف می‌شوند.

$$iff \begin{cases} \langle C, t_i \rangle \,\&\, \langle C, t_j \rangle \\ \langle C', t_i \rangle \,\&\, \langle C', t_k \rangle \end{cases} \rightarrow IS(t_i, t_k) \tag{۷}$$

این تعریف را می‌توان با استفاده از دو رابطه جایگزین‌پذیری مستقیم به شکل رابطه (۸) بازتعریف نمود.

$$iff\ \ DS(t_i, t_j) \,\&\, DS(t_i, t_k) \rightarrow IS(t_i, t_k) \tag{۸}$$

استفاده از تعریف جایگزین‌پذیری غیرمستقیم امکان برقراری ارتباط بین عبارات با مشابهت معنایی را در متن‌هایی با غنای زبانی کمتر فراهم می‌کند. در این صورت، حتی اگر یکی از عبارات مستقیماً در زمینه مشترک با دیگری قرار نگیرد، باز هم به عنوان کاندیدی برای برقراری رابطه مشابهت معنایی مورد توجه قرار می‌گیرد.

استفاده از این تعریف هرچند ممکن است به نظر مفید بیاید، امّا مشکلاتی را نیز فراهم می‌کند. به این معنا که ممکن است یک کلمه به دلیل اینکه تنها یک‌بار در یک پنجره معین واقع شده است، با یک عبارت مشابه فرض شود. چون عبارت واسطه نیز می‌تواند با کلمات دیگری جایگزین شود، این رابطه امکان برقراری مشابهت بین کلمه مذکور با تمام جایگزین‌های عبارت واسطه را فراهم می‌کند. برای کاستن از اثر جایگزین شدن یک عبارت با عبارت دیگر فقط به دلیل یک نوبت واقع شدن در یک زمینه مشترک، امتیاز شدّت جایگزین‌پذیری را تعریف می‌نماییم. این امتیاز براساس تعداد زمینه‌های مشترک بین دو عبارت تعریف می‌شود.

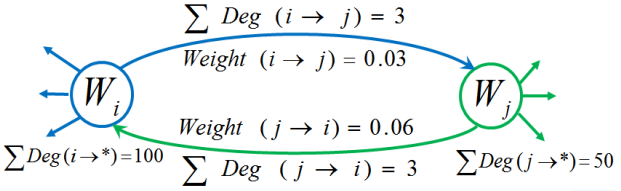
۳-۲- گراف جایگزین‌پذیری

برای تعریف امتیاز جایگزین‌پذیری باید به این نکته توجه نمود که این امتیاز باید بتواند میزانی برای شدّت رابطه بین دو کاندید جایگزینی را ارائه دهد. قبلاً روابط متفاوتی برای این منظور مورد استفاده قرار گرفته است که معروف‌ترین آن‌ها احتمال جایگزین شدن [۳۰] و آنتروپی جایگزینی [۳۱] دو کلمه با هم است. با انجام آزمایش‌ها مشخص گردید، استفاده از احتمال جایگزینی دو عبارت در مواردی که زمینه غیرمتداول تنها یک‌بار ظاهر شده باشد و دقیقاً در همان یک‌مرتبه نیز دو عبارت قابل جایگزین شدن باشند، عملی نیست. زیرا در این صورت احتمال جایگزینی برابر یک خواهد بود. چنین مواردی می‌تواند باعث جایگزین شدن یک عبارت پرت[۱۱] با یک عبارت واسطه شود. وجود یک واسطه متداول می‌تواند فرایند جایگزینی را به صورت زنجیروار گسترش دهد و امکان جایگزینی بسیاری از عبارات را با یک عبارت پرت فراهم نماید. در این صورت استفاده از مدل احتمال جایگزینی برای هر جفت نمی‌تواند مقیاس مناسبی از میزان مشابهت دو عبارتی به دست دهد که با تعداد کمی واسطه، باهم مشابه شده‌اند.

برای حل این مشکل از یک‌راه حل متداول در بازیابی اطلاعات، یعنی قدم زنی بر روی گراف روابط بین موجودیت‌ها بهره گرفته‌ایم. در این روش، به ازای هر عبارت کاندید، یک گره در گراف در نظر گرفته می‌شود و سپس به ازای هر زمینه مشترک بین دو کاندید، دو یال

جهت‌دار به گراف افزوده می‌شود. در این روش، ارتباطات مستقیم از طریق یال‌ها مدل می‌شوند و ارتباط غیرمستقیم از طریق دنبال‌کردن یال‌ها و تشکیل مسیر در گراف قابل استخراج است.

براساس این تعریف، به ازای هر یال خارج شده از یک گره، یک درجه به مجموع درجات خروجی آن افزوده می‌شود. بدیهی است ارزش وجودی تنها یک زمینه مشترک با تعداد تکرار ۱۰۰ مرتبه، بیش از وجود سه زمینه مشترک است که هر کدام تنها یک‌بار تکرار شده‌اند. بنابراین در محاسبه وزن یال‌های گراف، ارزش هر زمینه مشترک باید با توجه به تعداد موارد بروز آن تعیین شود و به‌جای ارزش خالص هر یال، از درجه نرمال شده بهره گرفته می‌شود. درجه نرمال هر یال از تقسیم درجه آن به مجموع درجات خروجی به دست می‌آید. در این حالت هر عبارت t_i به‌واسطه داشتن زمینه مشترک با عبارت t_j دارای یک یال خروجی به مقصد t_j و یک یال ورودی با مبدأ t_i است. امتیاز نرمال یال خروجی از t_j از تقسیم امتیاز این یال به درجات خروجی t_j به دست می‌آید و در روندی مشابه امتیاز یال خروجی t_i از تقسیم امتیاز آن بر مجموع درجات خروجی t_i. شکل شماره (۱) نمونه‌ای از این ارتباط را نشان می‌دهد.



شکل ۱: ارتباط بین دو گره در گراف ارتباط جایگزین پذیری را نشان می‌دهد.

۳-۳- امتیاز شباهت‌معنایی

گراف جایگزین‌پذیری درواقع نشان‌دهنده ساختار جایگزینی مندرج در متن پیکره موردنظر است. بدین معنا که بر اساس این گراف، می‌توان تعیین نمود هرکدام از عبارات با کدام عبارت دیگر در حداقل چند زمینه یکسان جایگزین شده‌اند. امکان طی مسیر بر روی گراف زمینه کاندیدشان جفت‌هایی را فراهم می‌آورد که به علت عدم غنای متن در یک زمینه واحدِ مشترک دیده نشده‌اند ولی بار معنایی مشابه داشته‌اند و در زمینه‌هایی مشترک با یک کلمه واحد دیگر اتفاق افتاده‌اند. طی مسیر در گراف می‌تواند برای زبان‌هایی مانند فارسی که ازنظر پیکره‌ای غنای قابل قیاس با زبان‌هایی مانند زبان انگلیسی را ندارند، مفهوم زمینه مشترک را تعمیم بدهد.

برای محاسبه امتیاز جایگزین‌پذیری هر دو جفت کلمه، از روالی مشابه با الگوریتم TrustRank بهره گرفته می‌شود که در ادامه توجیه مربوط به این انتخاب بیان می‌گردد. این الگوریتم توسط Gyöngyi و همکاران [۳۲] پیشنهاد شد و از آن برای کشف صفحات هرز در وب جهانی بهره گرفته می‌شود. این الگوریتم گراف پیوند بین صفحات وب را تشکیل داده و سپس صفحاتی را که با تأیید ناظر انسانی، قابل اعتماد تشخیص داده می‌شوند با امتیاز میزان اعتماد ۰/۵ شارژ می‌کند و سپس فرایند انتشار امتیاز اعتماد را به شیوه الگوریتم PageRank به انجام می‌رساند. در الگوریتم TrustRank میزان اعتماد کسب شده توسط هر گره، وابسته به تعداد یال‌های ورودی و همچنین طول مسیر مرتبط‌کننده آن گره به یکی از گره‌های هسته[۱۲] است. یعنی هر گره که تعداد بیشتری یال از گره‌های قابل اعتماد دریافت کند و یا در مسیر کوتاه‌تری بتوان از یکی از آنها به آن رسید، باید دارای اعتبار بیشتری باشد. در صورت نگاشت این اصل به رابطه جایگزین‌پذیری خواهیم دید که دو کلمه در صورتی‌که در تعدادی زمینه، با وزن بیشتر و یا در مسیر کوتاه‌تری از همدیگر قابل دستیابی باشند، دارای رابطه جایگزین‌پذیری قوی‌تری هستند و درنتیجه ارتباط معنایی بیشتری دارند. بر همین اساس، پس از تولید گراف یکی از کلمات به عنوان گره هسته انتخاب می‌شود و با امتیاز اولیه ۱ شارژ می‌گردد. سپس این امتیاز با رابطه (۹) در گراف به جریان می‌افتد.

$$r = \alpha\, T\, \vec{r} + (1-\alpha)\,\vec{d} \qquad (۹)$$

در این رابطه، T میزان اعتماد، α ضریب میرایی[۱۳] و d بردار توزیع یکنواخت امتیاز و r بردار امتیاز هر گره است. در روش محاسبه شباهت معنایی، به علت ثابت بودن هسته اولیه، این هسته با مقدار ۱ شارژ می‌شود. در ادامه، امتیاز هر گره با توجه به تمام گره‌های مرتبط با آن محاسبه می‌شود. استفاده از ضریب میرایی در این گراف توجیه پرش تصادفی (همانند تعریف مسئله وبگرد تصادفی در الگوریتم PageRank) را ندارد ولی به جهت حفظ شرط لازم برای همگرایی زنجیره‌های مارکف تشکیل شده بر روی گراف، در رابطه وارد می‌شود. باید توجه نمود که وجود این پرش در گراف جایگزینی به معنای پذیرش جایگزینی تصادفی بین هر دو کلمه ممکن است. بنابراین برای حفظ شرایط لازم برای همگرایی و درعین‌حال جلوگیری از ایجاد اختلال در روند محاسبه میزان شباهت معنایی، مقدار ضریب میرایی به عدد یک نزدیک خواهد بود تا احتمال پرش (α-۱) به صفر نزدیک شود.

۴-۳- پیاده سازی الگوریتم

قبل از تشریح مراحل پیاده‌سازی الگوریتم باید به تعریف زمینه در این تحقیق بپردازیم. در تحقیقات مختلف در حیطه پردازش‌های معنایی، تعاریف مختلفی برای زمینه ذکر شده است. پرکاربردترین تعریف در این حوزه، زمینه را به صورت مجموعه عبارات هم‌رخداد در یک پنجره از عبارات، به طول حداکثر k و در طرفین عبارت هدف، t_{target}، به شکل رابطه (۱۰) تعریف می‌نماید.

$$\langle t'_k t'_{k-1} \dots t'_1\ t_{target}\ t_1 \dots t_k \rangle \qquad (۱۰)$$

تمام کلمات موجود در عبارت‌های $t_1,t_2,\dots,t_k$ و $t'_1,t'_2,\dots,t'_k$ کلمات زمینه برای t_{target} نامیده می‌شوند. اجرای این روش نیازمند اجرای یک فرایند چهار مرحله‌ای است. برای اجرای این روش، یک پیکره زبانی که از نظر نویسه‌ای پالایش شده است انتخاب شده و مراحل زیر با استفاده از این دادگان به اجرا درمی‌آید:

۱) تولید دادگان چندجزیی

۲) تولید دادگان جایگزین‌پذیری

۳) تولید گراف جایگزین‌پذیری

۴) محاسبه امتیاز شباهت معنایی

در پیاده‌سازی این الگوریتم، طول پنجره محاسبه زمینه دقیقاً برابر یک واحد فرض می‌شود. بنابراین با فرض اینکه طول عبارت هدف برابر n باشد، باید پنجره‌ای به طول $n+2$ را از متن استخراج نمود. به همین دلیل برای تولید دادگان مربوط به زمینه می‌توان در ابتدا دادگان چندجزیی[۱۴] با طول دو واحد بیشتر از عبارات را از متن پیکره استخراج نمود. یعنی با داشتن رشته چندجزیی $\langle t_1\ t_2\ \dots\ t_{n+1}\ t_{n+2}\rangle$، می‌توان یک عبارت هدف به صورت $\langle t_2\ \dots\ t_{n+1}\rangle$ و یک زمینه به شکل $\langle t_1\ ,\ t_{n+2}\rangle$ استخراج نمود. برای جلوگیری از بروز خطای ناشی از وجود کلمات کم احتمال یا نادرست، تمام چند جزئی‌ها باید حداقل دارای تعداد تکرار مشخص باشند که برابر f_n تعیین می‌شود. در ادامه با فرض اینکه چند جزئی‌های موردنیاز تولید، شمارش و هرس شده‌اند، الگوریتم‌های لازم برای پیاده‌سازی این روش را توضیح خواهیم داد. اولین قدم در تولید جفت‌های جایگزین‌پذیر، تولید جفت‌هایی به صورت ⟨عبارت هدف ، زمینه⟩ است که این وظیفه توسط تابع K-Subs (مندرج در الگوریتم شماره (۱)) به انجام می‌رسد. این تابع با دریافت طول زمینه (در اینجا برابر یک) و یک چندجزیی، عبارت هدف را از زمینه جدا نموده و آنها را در قالب دوتایی‌های ⟨کلید ، زمینه⟩ تحویل می‌دهد. البته باید خاطرنشان نمود که چون داشتن زمینه‌های مشترک شرط جایگزینی جفت‌ها هستند، بنابراین کلید رابطه، "زمینه" و مقدار، "عبارت هدف" خواهد بود.

Algorithm 1: K-Subs generates the ⟨Target , Context⟩ pair from the Ngram g.

Input: Ngram $g = \langle w_1\ w_2\ \dots\ w_n\rangle$
Context Length k
Output: key-value pairs of target and context from Ngram g.

```
1  function K_Subs (g , k)
2     Key ← ⟨ w1 w2 ... wk   w(n-k+1) .... w(n-1) wn ⟩
3     Value ← w(k+1) ... w(n-k)
4  return ⟨Key , Value⟩
```

الگوریتم ۱: تابع تولیدکننده جفت‌های ⟨عبارت هدف ، زمینه⟩ از چند جزئی‌ها.

تابع نگاشت-کاهش (الگوریتم شماره (۲))، تابع K_Subs را فراخوانی می‌نماید و به‌واسطه آن برای تمام چند جزئی‌های موجود، زمینه و عبارات هدف را تعیین می‌نماید و سپس به ازای هر دو عبارت هدف ظاهر شده در یک زمینه یکسان، یک جفت به صورت $\langle\langle t_1, t_2\rangle,1\rangle$ می‌سازد. در این عبارت، دو عبارت هدف t_1 و t_2 ، کلید و مقدار رابطه برابر ۱ است. به بیان بهتر، هر دو جفت از عبارت‌های هدف مربوط به یک کلید، به ازای هر بار دیده‌شدن در یک زمینه مشترک یک‌بار شمارش می‌شوند. در قدم بعدی، جفت‌های دیده شده باید شمارش شوند که این وظیفه توسط تابع نگاشت-کاهش مندرج در الگوریتم (۳) به انجام می‌رسد.

این تابع جفت عبارت‌های قابل جایگزینی را به همراه تعداد تکرار آنها به عنوان خروجی تولید می‌کند. این جفت‌ها دارای قالبی به صورت $\langle\langle t_1, t_2\rangle, f\rangle$ هستند. در ادامه با استفاده از این جفت‌ها که وجود یال بین گره‌های گراف جایگزینی را تعیین می‌کنند، گراف جایگزینی تولید می‌شود. هر یال از این گراف با استفاده از تعداد تکرار آن جفت که با تقسیم بر مجموع درجات خروجی یک گره نرمال شده است، وزن‌دهی می‌شود. باید خاطرنشان کرد که به ازای هر جفت قابل جایگزینی، دو یال در دو جهت مختلف با وزن‌های متفاوت به گراف افزوده می‌شود. پس از آماده‌سازی گراف، می‌توان از آن برای محاسبه شباهت معنایی هر دو جفت کلمه بهره گرفت.

Algorithm 2: MapReduce substitutable pair generator.

Input: Ngram Set G_n
Maximum Context Length l
Output: Substitutable pairs.

```
1  function map (Gn , k)
2     foreach g in set Gn do
3        for k ←1 to l do
4           ⟨Key , Value⟩= K-Subs (g , k);
5           emit (Key , Value);
6        end
7     end
8  end
9  function reduce (Key, List ⟨value⟩ )
10    foreach Key do
11       foreach v1 in List ⟨value⟩ do
12          foreach v2 in List ⟨value⟩ do
13             if v1<v2 then
14                emit (⟨v1 v2⟩ , 1 );
15             end
16          end
17       end
18    end
19 end
```

الگوریتم ۲: تابع نگاشت-کاهش تولید کننده‌ی جفت‌های جایگزین‌شونده با استفاده از الگوریتم شماره ۱.

Algorithm 3: MapReduce substitutable pair counting.

Input: Substitutable pairs.
Output: Substitutable pairs and their counts.

```
1  function map (Key, Value)
2     emit (Key , Value);
3  end
4  function reduce (Key, List ⟨value⟩ )
5     int freq← 0
6     foreach Key do
7        foreach v in List ⟨value⟩ do
8           freq ← freq+1;
9        end
10       emit (Key , freq);
11    end
12 end
```

الگوریتم ۳: تابع نگاشت-کاهش شمارش جفت‌های جایگزین‌پذیر به‌دست‌آمده از الگوریتم شماره ۲.

فرایند تولید گراف با استفاده از جفت‌های قابل جایگزینی و تعداد آنها در الگوریتم شماره (۴) قابل مشاهده است. خطوط ۲ تا ۷ با دریافت جفت‌های مذکور، یال‌هایی را در هر دو جهت برای گره‌ها تعریف می‌کنند که در محاسبه وزن آنها از تعداد تکرار آن جفت

استفاده می‌نمایند. در این الگوریتم، برای همه گره‌ها مقدار امتیاز اولیه ثابت تعیین شده و به‌طور اخص برای گره مربوط به عبارت هدف، امتیاز ۱ به‌علاوه امتیاز اولیه لحاظ می‌شود. در خطوط ۱۳ تا ۲۱ این الگوریتم، امتیاز هر یک از گره‌ها به صورت تکراری محاسبه می‌شود. محاسبه امتیاز در این روش مشابه الگوریتم PageRank است. برای این منظور امتیاز یک گره با استفاده از مجموع امتیازات گره‌های منتهی به آن محاسبه می‌شود (ر.ک. خط ۱۶). تعداد مرتبه‌های تکرار این الگوریتم برابر قطر[۱۵] گراف خواهد بود.

Algorithm 4: Substitution score propagation algorithm.

Input: Substitutable pair set, E
Seed and Sink words, i.e. w_{seed} , w_{sink}
Size of Vocabulary, N
Damping factor, d
Graph diameter, *diameter*
Output: Score[w_{sink}]

```
1   function GraphScore (.)
2       foreach e =⟨⟨w_i ,w_j⟩ , f ⟩ in E do
3           outdeg[w_i]← outdeg[w_i]+f ;
4           outdeg[w_j]← outdeg[w_j]+f ;
5           w_i.Neighbors.add(w_j , f );
6           w_j.Neighbors.add(w_i , f );
7       end
8       foreach w_i in Vocab do
9           Sim[w_i]← 1/N;
10      end
11      Sim[w_seed]← Sim[w_seed]+1;
12      it ← 1;
13      while it<diameter do
14          foreach w_i in Vocab do
15              foreach n in w_i.Neighbors do
16                  Sim[w_i]← Sim[w_i]+
                        d*(Sim[n]/outdeg[n]);
17              end
18              Sim[w_i]← Sim[w_i]+ (1-d ) / N ;
19          end
20          it← it-1;
21      end
22  return Sim[w_sink];
```

الگوریتم ۴: تابع محاسبه امتیاز گره دریافت کننده با استفاده از جفت‌های جایگزین‌پذیر و گره هسته.

برای محاسبه امتیاز جایگزینی یک جفت کلمه، ابتدا کلمه اوّل به عنوان گره هدف انتخاب شده و امتیاز دریافتی از آن توسط گره دوم (گره دریافت‌کننده[۱۶]) محاسبه می‌شود و در مرحله دوم این روند برعکس شده و امتیاز دریافتی گره اوّل از گره دوم محاسبه شده و میانگین این دو امتیاز به عنوان میزان شباهت معنایی مورد استفاده قرار می‌گیرد. مراحل اجرایی فرایند تعویض گره‌های هدف و دریافت‌کننده و میانگین‌گیری برای محاسبه امتیاز مشابهت معنایی در الگوریتم شماره (۵) قابل مشاهده است. در واقع این تابع اصلی[۱۷] است که باید برای هر جفت کلمه فراخوانی شود و به دنبال فراخوانی آن، تابع (۴) نیز فراخوانی می‌شود.

۴- آزمایش‌ها و بررسی نتایج

فرایند بررسی میزان کارایی روش پیشنهادی در مقایسه با روش‌های موجود نیازمند مراحلی از قبیل آماده‌سازی دادگان آموزش و آزمون، تعریف معیارهای ارزیابی و پیاده‌سازی روش‌های مشابه است که در این بخش به تفضیل بیان خواهند شد.

Algorithm 5: Semantic Similarity Estimation (SSE).

Input: Substitutable pair set, E
Seed and Sink words, i.e. w_{seed} , w_{sink}
Size of Vocabulary, N
Graph diameter, *diameter*
Output: *Semantic_Sim*(w_{seed} , w_{sink})

```
1   function SSE (.)
2       d←1-ε ;
3       S_sink← GraphScore(E , w_seed ,w_sink ,N,d,diameter);
4       S_seed← GraphScore(E , w_sink ,w_seed ,N,d,diameter);
5       Semantic_Similarity←0.5*(S_seed+S_sink);
6   return Semantic_Similarity;
```

الگوریتم ۵: تابع محاسبه شباهت معنایی دو کلمه.

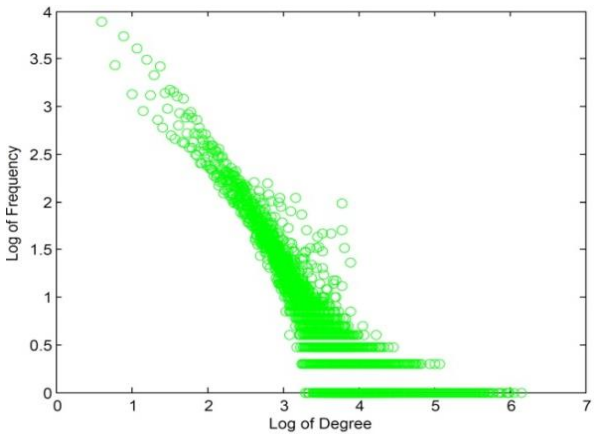
۴-۱- آماده‌سازی دادگان

برای اجرای روش پیشنهادی و روش‌های انتخاب شده برای مقایسه، دو دسته داده آماده شده است که عبارت‌اند از:

۱- دادگان آموزش

۲- دادگان آزمون

دادگان آموزشی که برای تولید گراف جایگزینی، مورد استفاده قرارگرفته است مجموعه ویکی‌پدیای فارسی بوده است. برای این منظور نسخه کامل همه اسناد ویکی‌پدیا مربوط به تاریخ یکم می ۲۰۱۶ مورد استفاده قرارگرفته است. این مجموعه شامل ۲۳۶۲۴۱۷ سند است و نیازمند فضایی به اندازه ۳/۸ گیگابایت برای ذخیره‌سازی است. مجموعه این اسناد در قالب یک فایل یکپارچه *XML* ذخیره می‌شوند که پس از تجزیه و تبدیل به قالب مسطح قابل استفاده است. یکی از مهم‌ترین چالش‌های دادگان زبانی در فارسی، تنوع نویسه‌های[۱۸] مورد استفاده توسط کاربران است. باتوجه به مشابهت نویسه‌های فارسی و عربی، در بسیاری موارد، کاربران از نویسه‌های عربی در اسناد فارسی استفاده می‌کنند. برای جلوگیری از بروز هرگونه خطا، تمام نویسه‌های مورد استفاده در این دادگان به نویسه‌های فارسی تغییر داده شده است. پس از یکسان‌سازی نویسه‌ها، برچسب‌های دستوری مورد استفاده توسط بنیاد ویکی‌پدیا که برای ساختارده به اسناد، تولید جداول، تغییر مسیر و اعلام وضعیت بکار گرفته می‌شوند از اسناد حذف شدند. اسناد براساس نقطه‌گذاری استاندارد فارسی و همچنین با لحاظ تقسیم‌بندی‌های ساختاری ویکی‌پدیا قسمت‌بندی شده‌اند. به‌جز پردازش‌های ذکر شده پردازش دیگری برای آماده‌سازی دادگان مورد نیاز نیست و حتی حذف کلمات توقف نیز برای این دادگان صورت نمی‌گیرد.



شکل ۲: نمودار log-log توزیع درجات گراف جایگزینی.

با توجه به اینکه دادگان مورد استفاده در مرحله آزمون دارای عباراتی با طول ۱ و ۲ کلمه بودند، برای تولید جفت‌های جایگزینی، تمام سه‌جزئی‌ها[۱۹] و چهار جزئی‌های[۲۰] متن استخراج شده‌اند. به این ترتیب در هر طرف از یک عبارت یک کلمه‌ای و دوکلمه‌ای دقیقاً یک عبارت یک کلمه‌ای به عنوان زمینه قرار می‌گیرد. برای تولید زمینه مربوط به هر عبارت، با انتخاب هر چندجزیی، کلمات ابتدایی و انتهایی آن به عنوان زمینه انتخاب شده و یک یا دوکلمه وسط به عنوان عبارت هدف در نظر گرفته شده‌اند. پس از شمارش این چند جزئی‌ها، مواردی که دارای حداقل تعداد تکرار ۲ واحد بوده‌اند از این مجموعه استخراج شده و سایر موارد حذف شده‌اند.

با توجه به حجم دادگان، برای تولید این مجموعه از یک برنامه در چهارچوب نگاشت-کاهش[۲۱] بر روی بستر محاسبات توزیع شده هدوپ[۲۲] بهره‌گرفته شده است. پیش‌تر به الگوریتم‌های (۱) تا (۳) اشاره نمودیم که مراحل فرایند تولید و شمارش جفت‌های جایگزینی را به انجام می‌رسانند. پس از تولید جفت‌ها، با تعیین حداقل تعداد تکرار ۴ برای خروجی، ۲۶۹۱۴۵۶۸ جفت جایگزینی در این مرحله به دست می‌آید. جفت‌های مذکور، دارای مینیمم تعداد جایگزینی ۴ و ماکزیمم تعداد جایگزینی ۳۰۰ بوده‌اند. از مجموعه این جفت‌ها گرافی با ۴۸۴۸۲۵ گره و ۲۶۹۱۴۵۶۸ یال و با قطر ۶ تشکیل شده‌است. بنابراین حداکثر تعداد اجرای لازم برای اطمینان از دریافت امتیاز توسط هر گره دریافت‌کننده ۶ مرحله است. این گراف با توجه به‌اندازه و نوع توزیع درجات خروجی و ورودی دارای ساختار یک "دنیای کوچک"[۲۳] است. نمودار لگاریتم-لگاریتم توزیع درجات خروجی این گراف در شکل شماره (۲) قابل ملاحظه است.

برای آزمون نتایج به‌دست‌آمده از این روش، از یکی از مجموعه دادگان آزمون با شهرت جهانی در این عرصه بهره‌گرفته شده است. دادگان RG-65 اولین بار توسط Rubenstein و Goodenough [۳۳] برای بررسی میزان کارایی روش‌های تعیین مشابهت معنایی در زبان انگلیسی آماده شده و مورد استفاده بسیاری از محققین قرارگرفته‌است. از آنجا که این دادگان تنها برای زبان انگلیسی قابل استفاده بودند، تیم‌های متعددی در سراسر جهان، با رعایت ضوابط تعیین شده در جریان تولید دادگان RG-65 به زبان انگلیسی، نمونه‌های آن در زبان‌های دیگر را نیز آماده نموده‌اند. در این راستا Camacho-Collados و همکاران [۳۴] نسخه فارسی این دادگان را با ضوابط مشابه برای زبان فارسی تولید نموده‌اند و از سال ۲۰۱۵ با انتشار رسمی این داده، از آن به عنوان مرجع مورد استفاده برای تعیین مشابهت معنایی کلمات در زبان فارسی بهره گرفته می‌شود.

این دادگان جز یکسان سازی نویسه‌ای برای حرف "ی"، نیاز به پردازش دیگری ندارد. باید خاطرنشان نمود که دادگان آموزش و آزمون برای همه روش‌ها یکسان بوده و هیچ پیش‌پردازش خاصی برای یکی از روش‌ها صورت نگرفته است که برای سایرین صورت نگرفته باشد.

۴-۲- روش‌های مقایسه شده با روش پیشنهادی

برای اجرای آزمایش‌های مورد نیاز برای ارزیابی این روش از دسته‌های مختلف از روش‌های بدون سرپرست، یک نماینده انتخاب شده و برای مقایسه با این روش پیاده‌سازی شده‌اند. صرف‌نظر از تفاوت‌های این روش‌ها، همه آنها با استفاده از دادگان آموزش و آزمون مشابه اجرا شده‌اند. از دسته روش‌های استفاده کننده از معیارهای مشابهت معنایی و زمینه مشترک یک روش انتخاب شده است. این روش اولین بار توسط InkPen [۳۵] پیشنهاد شد و از PMI مابین هر کاندید جایگزینی با تمام عبارات موجود در مجموعه زمینه مشترک بهره می‌گیرد. جزئیات این روش در مقاله [۳۵] قابل ملاحظه است.

از دسته روش‌های مبتنی بر بردار، روش GLOVE [۲۵] و از دسته روش‌های مبتنی بر احتمالِ زمینه، روش پیشنهادی توسط Xue و همکاران در [۳۶] مورد استفاده قرار گرفته است. دلیل انتخاب روش‌های مندرج در مقالات [۳۵]، [۳۶] استفاده هر دو تحقیق از زمینه مستخرج از چند جزئی‌ها و موفقیت چشمگیر آنها بوده است؛ امّا روش GLOVE [۲۵] صرفاً به علت اینکه موفق‌ترین روش در دسته روش‌های تعیین مشابهت معنایی با استفاده از بردارهای معنایی بوده، انتخاب شده است. باید خاطرنشان نمود که این روش نیز از اطلاعات زمینه عبارات در قالب هم‌رخدادی بهره می‌برد ولی به هیچ وجه از مدل چندجزیی احتمالی برای محاسبات خود استفاده نمی‌کند. فرایند آموزش GLOVE با احتساب اندازه ۵۰۰ بعد برای بردارهای توصیف‌کننده کلمات، برای تولید امتیاز مشابهت معنایی بر روی یک بستر سخت‌افزاری با ۲۴ پردازشگر(CPU) و GB۱۲۰ حافظه به صورت چندنخی[۲۴] اجرا شده است.

۴-۳- معیارهای ارزیابی

با توجه به اینکه میزان مشابهت معنایی یک مفهوم کیفی است، برای کمّی‌سازی آن از نظرات کاربران انسانی بهره گرفته می‌شود. یکی از متداول‌ترین روش‌ها در این زمینه، نمایش جفت‌هایی از کلمات برای کاربران و درخواست از آنها برای تعیین مشابهت معنایی است. در این

روش، کاربران یک امتیاز گسسته در بازه تعیین شده را به هر جفت اختصاص می‌دهند و در نهایت با استفاده از تکنیک‌های آماری یک مقدار معین از آرای کاربران استخراج و به عنوان میزان مشابهت معنایی آن دو جفت کلمه لحاظ می‌شود. از آنجا که این مقادیر در بازه دلخواه تغییر می‌کنند و چندان دقیق نیستند برای ارزیابی میزان دقت روش‌های تخمین مشابهت معنایی از ضریب همبستگی *Pearson* و یا معیارهای مشابه بهره‌گرفته نمی‌شود.

معیار ارزیابی در آزمون‌های مشابهت معنایی اصولاً بر توانایی روش‌ها بر ارائه لیستی با ترتیب مشابه با دادگان آزمون تاکید دارند و به همین دلیل از معیارهایی مانند ضریب همبستگی Spearman موسوم به ضریب Spearman's ρ و همچنین ضریب Kendall موسوم به Kendall's τ بهره گرفته‌می‌شود. این ضرایب به صورت مندرج در روابط (۱۱) و (۱۲) محاسبه می‌شوند.

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d(r_i - r_i')^2}{n(n^2-1)} \qquad (۱۱)$$

در رابطه (۱۱) فرایند محاسبه ضریب همبستگی *Spearman* نشان داده شده است. در این رابطه n تعداد عناصر موجود در لیست‌های مرتب شده، r_i و r'_i رتبه دو عنصر در دو لیست و تابع $d(.)$ میزان اختلاف دو رتبه مربوط به دو عنصر است. در واقع مجموع مربعات اختلاف رتبه‌های عناصر نظیر در این رابطه مورد استفاده قرار می‌گیرد. رابطه (۱۲) نحوه محاسبه ضریب *Kendall* را نشان می‌دهد.

$$\tau = 1 - \frac{\Delta(r_i, r_j)}{\frac{1}{2}n(n-1)} \qquad (۱۲)$$

در رابطه (۱۲)، تابع $\Delta(.,.)$ تعداد مواردی را تعیین می‌کند که رتبه بین هر دو جفت ممکن مانند r_i و r_j با هم یکسان نباشد. در واقع این معیار کسر تعداد جفت‌های با رتبه یکسان بین هر دو لیست را محاسبه می‌کند.

۴-۴- تحلیل نتایج آزمایش‌ها

نتایج ارزیابی روش پیشنهادی با سه روش معرفی شده با توجه به دو معیار ارزیابی فوق‌الذکر در جدول شماره (۱) قابل مشاهده است. نتایج آزمایش‌ها با استفاده از دادگان آزمایش نشان می‌دهد، روش پیشنهادی در ارزیابی با استفاده از ضریب ρ نتایج بهتری نسبت به روش‌های دیگر نشان داده است.

جدول ۱: نتایج مقایسه روش پیشنهادی با سایر روش‌ها.

#	الگوریتم	Spearman's ρ	Kendall's τ
۱	الگوریتم GlOVE [۲۵]	۰/۲۴۳۳	۰/۱۷۹۲
۲	الگوریتم Inkpen [۳۵]	۰/۳۲۱۷	۰/۲۲۶۱
۳	الگوریتم Xue و همکاران [۳۶]	۰/۱۹۲۹	۰/۱۲۵۲
۴	الگوریتم پیشنهادی	۰/۳۵۳۹	۰/۲۵۳۲

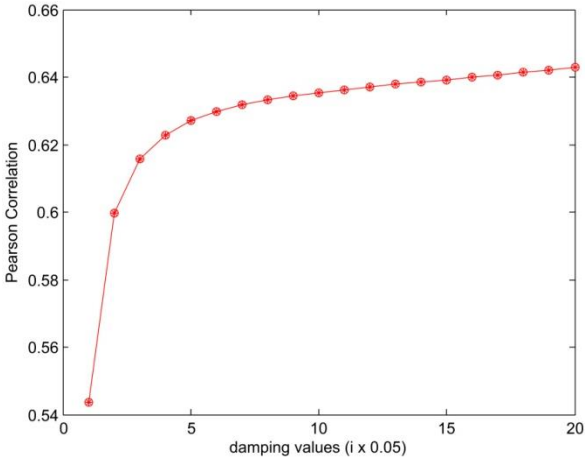
برتری این روش بر روشی مانند GLOVE که از روش‌های موفق برای تعیین مشابهت معنایی توزیع شده است امتیاز مهمی برای این روش محسوب می‌شود. این دستاورد، مخصوصاً با توجه به این نکته که فرایند آموزش الگوریتم GLOVE نیازمند حجم عظیم از محاسبات ماتریسی و برداری برای تولید بردارهای کلمات است، مسئله مهمی قلمداد خواهد شد. اما در ارزیابی با استفاده از *Kendall's τ* مشخص گردید، روش پیشنهادی کارایی کمتری داشته است. دلیل این امر را باید در تفاوت ساختاری نحوه محاسبه این دو ضریب جست. از آنجا که ضریب *Kendall* برای محاسبه تنها به ترتیب تمام جفت‌های موجود در لیست‌ها توجه می‌کند، به‌واسطه تغییرات کوچک در ترتیب در لیست‌ها این مقدار می‌تواند تغییرات قابل توجهی داشته باشد. امّا *Spearman's ρ* از میزان اختلاف رتبه عناصر در لیست‌های مرتب استفاده می‌نماید و درنتیجه در صورتی که عناصر لیست با ترتیب کمی متفاوت به دلیل تغییرات مقداری ناچیز قرار گرفته باشند، ناچیز بودن تغییرات، صرف‌نظر از تغییر ترتیب، مانع از بروز اختلاف در مقدار خروجی این پارامتر می‌شود. اصولاً در مواردی که مقادیر خروجی روش‌های دارای بازه تغییرات محدود و کوچک باشند، اثر پارامتر *Spearman's ρ* محسوس‌تر خواهد بود و خطای کمتری را به‌واسطه جابجای در لیست‌های مرتب که ناشی از نویز حاصل از اختلاف ناچیز مقادیر در مسئله هستند را متوجه نتایج می‌کند.

مسئله دیگری که در حین آزمایش‌ها باید به آن توجه نمود، تعیین میزان ضریب میرایی مناسب برای آزمایش‌ها است. چنانچه در بخش سوم نیز بیان شد، با توجه به ماهیت مسئله، حداکثر ضریب میرایی به معنای حداقل احتمال پرش تصادفی بین گره‌ها در گراف جایگزین‌پذیری است. ازآنجاکه از دید نظری، پرش تصادفی مخل تعاریف مشابهت معنایی مبتنی بر جایگزین پذیری است، بنابراین به نظر می‌رسد بهترین نتایج در چهارچوب پردازشی معرفی شده با بیشترین مقادیر ضریب میرایی به دست آید. در این راستا و به جهت اطمینان از صحت این فرضیه، یک آزمایش دیگر نیز به اجرا درآمده است.

از آنجا که مقادیر مشابهت معنایی در دادگان RG-65 در بازه [۴،۰] تغییر می‌کنند، برای هر محدوده به طول یک واحد یک نماینده انتخاب شده (در مجموع ۵ نماینده) و فرایند محاسبه مشابهت معنایی برای این پنج نماینده با مقادیر مختلف ضریب میرایی به اجرا در آمده است. مقادیر ضریب میرایی در این آزمایش‌ها در بازه [۱/۰ ، ۰/۰] با گام ۰/۰۵ تعیین شده و در نتیجه برای هر یک از این جفت‌ها، فرایند محاسبه در ۲۰ نوبت تکرار شده است. به صورت استثنا برای حفظ شرط همبستگی آخرین آزمون به‌جای مقدار ۱ با ε-۱ [۲۵] مقدار گرفته است تا شرط همگرایی لحاظ شود.

امتیازات کسب شده در این مراحل برای هر یک از نمونه‌ها ثبت و در نهایت با مقادیر مندرج در دادگان مقایسه شده‌اند. از آنجا که ترتیب مقادیر بین جفت‌های مختلف در همه نمونه‌ها یکسان بوده است، میزان

تغییرات مقادیر خالص مورد توجه قرار گرفته به همین دلیل از پارامتر ضریب همبستگی پیرسون برای مقایسه مقادیر به‌دست‌آمده در آزمایش‌ها با مقادیر درج شده در دادگان بهره گرفته شده است. نتایج این آزمایش‌ها در نمودار شکل شماره (۳) قابل ملاحظه است. چنانکه از شکل نیز برداشت می‌شود، ضریب میرایی پس از عبور از مقدار ۰/۸ تنها ۰/۰۰۲۲ واحد در مقدار ضریب همبستگی پیرسون مابین نمونه‌ها و دادگان آزمون تغییر ایجاد نموده است، امّا کماکان بالاترین میزان ضریب همبستگی به مقدار ε-۱ اختصاص دارد. به همین دلیل در طول آزمایش‌ها با توجه به سایر مقادیر و دقت تعیین شده، ضریب میرایی با مقدار ۰/۹۹ تنظیم شده است تا هم حداکثر مقدار ممکن را اتخاذ نموده باشد و هم شرط همگرایی لازم برای محاسبه امتیازات را ارضا نماید.



شکل ۳: تغییرات همبستگی بین مقادیر با تغییرات ضریب میرایی.

نتایج آزمایش‌ها نشان دهنده این واقعیت است که نتایج به دست آمده از روش پیشنهادی بسیار شبیه به نتایج مقاله [۳۵] است، همچنین این دو روش به لحاظ ساختاری بسیار شبیه به هم هستند و ممکن است این پرسش به وجود آید که در این صورت روش پیشنهادی چه برتری نسبت به روش پیشنهادی Inkpen [۳۵] دارد؟

جدول شماره (۲): نتایج مقایسه زمانی روش پیشنهادی با سایر روش‌ها.

#	الگوریتم	زمان پیش‌پردازش	زمان آموزش	زمان محاسبه مشابهت
۱	الگوریتم GlOVE [۲۵]	۱۱ دقیقه	۳۲ ساعت	۴ میلی ثانیه
۲	الگوریتم Inkpen [۳۵]	۱۱ دقیقه	-	۲۳۲۴ میلی ثانیه
۳	الگوریتم Xue و همکاران [۳۶]	۱۱ دقیقه	-	۳۱۵۳ میلی ثانیه
۴	الگوریتم پیشنهادی	۱۴ دقیقه	-	۲۴۳ میلی ثانیه

در پاسخ به این سؤال باید گفت، مهم‌ترین برتری روش پیشنهادی به برتری زمانی این روش مربوط می‌شود. از بین این روش‌ها، روش‌های Xue، Inkpen و روش پیشنهادی تنها به یک مرحله پیش‌پردازش داده‌ای نیاز دارند ولی روش GLOVE نیازمند فرایند آموزش است. از طرفی هر یک از این روش‌ها با دسترسی به دادگان آماده شده به زمانی برای تخمین مشابهت معنایی نیازمند هستند. در این راستا، زمان‌های لازم برای تولید دادگان اولیه، زمان آموزش و متوسط زمان محاسبه مشابهت معنایی را برای هر سه روش مقایسه خواهیم نمود. باید خاطرنشان نمود که این زمان از متوسط‌گیری بر روی زمان‌های صرف شده برای تمام جفت‌های موجود در دادگان آزمون به دست آمده است. همچنین باید به این نکته اشاره نمود که همه روش‌های فوق یک‌بار نیازمند خواندن همه دادگان هستند و در مراحل بعدی تنها باید بر روی دادگان موجود در حافظه فرایند جستجو را به اجرا گذارند، بنابراین از این زمان برای همه روش‌ها صرف‌نظر شده است. از طرفی فرایند محاسبه امتیاز انتشار در گراف جایگزینی را نیز می‌توان با استفاده از روش‌های ابتکاری بهینه نمود و از پیمایش مسیرهای اضافی خودداری نمود به همین دلیل مقدار پیمایش گراف فوق که یک گراف تنک نیز هست بسیار کم خواهد بود. در هنگام مقایسه این روش‌ها، ممکن است این پرسش مطرح شود که چرا روش GLOVE نباید به‌جای الگوریتم پیشنهادی انتخاب شود؟

پاسخ این پرسش در فرایند به‌روزرسانی مجموعه دادگان نهفته است. جدول شماره (۳) زمان لازم برای افزودن تنها یک کلمه را به هر یک از مدل‌ها نشان می‌دهد. چنانکه از این جدول برمی‌آید برای درج تنها یک کلمه در مجموعه کلمات دادگان الگوریتم GLOVE باید فرایند پرهزینه بازآموزی را به انجام رساند، درحالی‌که برای درج همان کلمه در دادگان الگوریتم پیشنهادی زمان بسیار کمتری موردنیاز خواهد بود.

جدول ۳: نتایج مقایسه زمان لازم برای افزودن یک کلمه جدید به مدل‌های هر یک از روش‌ها.

#	الگوریتم	زمان پیش-پردازش	زمان آموزش
۱	الگوریتم GlOVE [۲۵]	۳۶۴ ثانیه	۵۲ دقیقه
۲	الگوریتم Inkpen [۳۵]	۳۶۴ ثانیه	-
۳	الگوریتم Xue و همکاران [۳۶]	۳۶۴ ثانیه	-
۴	الگوریتم پیشنهادی	۴۳۰ ثانیه	-

۵- نتیجه‌گیری

در این مقاله، یک روش جدید برای محاسبه میزان مشابهت معنایی در کاربردهای مختلف پردازش زبان پیشنهاد شده است. این روش از تعریف وجود زمینه مشترک بین کلمات بهره گرفته است. تفاوت عمده رویکرد اتخاذ شده در این تحقیق با تحقیقات مشابهِ مبتنی بر استفاده از زمینه مشترک، تعریف امکان جایگزین‌پذیری غیرمستقیم است. این نوع از جایگزین‌پذیری براساس رابطه‌ای تعریف می‌شود که به‌واسطه جایگزین‌پذیری یک واسطه مشترکِ ثالث بین دو عبارت به دست می‌آید. مهم‌ترین ویژگی مثبت این روش، امکان ایجاد رابطه بین جفت‌هایی است که به‌واسطه غنای کم پیکره در زمینه مشترک بروز ننموده‌اند ولی قابلیت جایگزین شدن باهم را دارند.

ممکن است مشابهتی بین این روش و مجموعه روش‌هایی که از قدم‌زنی تصادفی استفاده نموده‌اند به نظر برسد ولی براساس اطلاع نویسندگان این مقاله، هنوز روشی که از قدم‌زنی تصادفی بر روی گراف جایگزین‌پذیری یا هر گراف تولید شده در یک فرایند بدون‌سرپرست استفاده کند، پیشنهاد نشده است و اکثر روش‌هایی که از الگوریتم PageRank و یا موارد مشابه بهره برده‌اند، از دادگان مستخرج از هستان‌شناسی‌ها، پایگاه‌های دانش و یا مفاهیم مستخرج از دائره‌المعارف‌های تحت وب برای تولید گراف انتشار مشابهت معنایی بهره برده‌اند. استفاده از گراف‌های این چنینی گرچه به نظر دقیق می‌رسد ولی مستلزم انجام پردازش‌های حجیم و زمان‌بر است. درحالی‌که گراف جایگزینی فوق، علی‌رغم اندازه بزرگ به‌راحتی بر روی بسترهای توزیع شده متداول قابل تولید است. یکی از مسائلی که این تحقیق را نسبت به سایر تحقیقات همنوع خود که از اطلاعات زمینه مشترک (مانند [۳۱] و [۳۷- ۳۵]) بهره گرفته‌اند، متمایز می‌کند؛ توانایی این روش در ارائه دقتی مناسب بر روی پیکره‌ای کوچک‌تر از پیکره مورد استفاده در تحقیقات مشابه است. در تحقیق [۳۵] از پیکره چندجزیی گوگل بهره گرفته شده است و یا در تحقیقاتی مانند [۳۸]، [۳۹] از پیکره ویکی‌پدیای انگلیسی؛ درحالی‌که در این تحقیق روش‌های فوق بر روی پیکره ویکی‌پدیای فارسی اجرا شده است که از نظر تعداد اسناد و حجم متن (بر حسب تعداد توکن‌ها) به مراتب کوچک‌تر از ویکی‌پدیای انگلیسی است.

از بین روش‌های مورد ارزیابی در مقایسه این روش، روش GLOVE که دارای الگوریتمی مشابه الگوریتم Word2Vec است و در بسیاری از ارزیابی‌ها توانسته است از الگوریتم Word2Vec بهتر عمل کند، می‌تواند در نقاط تنک از دامنه مسئله بر روش پیشنهادی غلبه کند. با توجه به اینکه روش پیشنهادی به حداقل دو مورد مشابهت دامنه برای تعیین رابطه جایگزینی احتیاج دارد، در تعیین مشابهت معنایی در مورد کلماتی با تعداد تکرار بسیار پایین (مثلاً در ویکی‌پدیای فارسی، کلمات با تعداد تکرار زیر ۵ دارای تکرار بسیارپایین قلمداد می‌شوند) که به ندرت در حداقل دو زمینه یکسان قرار می‌گیرند، چندان کاربردی ندارد. در واقع روش پیشنهادی در برخورد با کلمات با تعداد تکرار بالا یا متوسط عملکردی بهتر از GLOVE ارائه می‌کند ولی درمورد کلماتی با احتمال وقوع بسیار کم، معمولاً GLOVE با توجه به غیرمتقارن درنظرگرفتن زمینه می‌تواند بهتر عمل کند. از آنجا که تعیین مشابهت معنایی برای چنین کلمات کم‌کاربردی چندان استفاده‌ای ندارد، بنابراین عدم توانایی الگوریتم پیشنهادی برای تعیین مشابهت معنایی این کلمات اثری بر کارایی کلی آن ندارد.

مقایسه زمان لازم برای آماده‌سازی دادگان و آموزش نشان می‌دهد روش پیشنهادی و روش Inkpen [۳۵] بسیار شبیه به هم هستند. این دو روش می‌توانند در معیارهای همبستگی کارایی بسیار نزدیکی را ارائه دهند ولی روش پیشنهادی در زمان اجرا در حدود ۱۰ مرتبه سریع‌تر است. در مقایسه زمان اجرای هر یک از روش‌ها، روش GLOVE [۲۵] نیازمند زمان بسیار کمتری نسبت به سایرین است امّا این روش برای هر به‌روزرسانی دادگان نیاز به بازآموزی مجدد دارد که به‌مراتب طولانی‌تر از زمان لازم برای سایر روش‌ها است.

نویسندگان این مقاله در نظر دارند در راستای ادامه این تحقیق، با توجه به کارایی این روش در تعیین مشابهت معنایی کلمات از آن برای تولید کاندیداهای فرایند استخراج بدون سرپرست کلمات مترادف استفاده نمایند و با استفاده از ترکیب آن با ویژگی‌های آماری دیگر قابل استخراج از متن پیکره به امتیازدهی انواع عبارت‌های مترادف بپردازند.

سپاسگزاری

این مقاله با پشتیبانی مالی و علمی موتور جستجوی پارسی‌جو به انجام رسیده است و تمام محاسبات مربوط به پردازش دادگان حجیم در این مقاله بر روی بستر نگاشت-کاهش این پروژه به اجرا درآمده است. نویسندگان مراتب تشکر خود را از اعضای تیم پارسی‌جو و به‌طور اخص خانم‌ها صدیقه طباطبایی و مهدیه فلاح و آقای امین رئیس‌زاده ابراز می‌دارند. همچنین باید خاطرنشان نمود که پیاده‌سازی مقیاس بزرگ مدل‌های مبتنی بر شبکه عصبی بدون کمک آقای محمدصادق طاهرزاده عملی نبود.

مراجع

[1] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller., “Introduction to WordNet: an on-line lexical database,” *Int. J. Lexicogr.*, vol. 3, no. 4, pp. 235–244, 1990.

[2] R. Rapp, “The automatic generation of thesauri of related words for English, French, German, and Russian,” *Int. J. Speech Technol.*, vol. 11, no. 3, pp. 147–156, 2009.

[3] F. Diaz, B. Mitra, and N. Craswell, “Query Expansion with Locally-Trained Word Embeddings,” in *The 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016.

[4] G. Zuccon, B. Koopman, P. Bruza, and L. Azzopardi, “Integrating and Evaluating Neural Word Embeddings in Information Retrieval,” in *Proceedings of the 20th Australasian Document Computing Symposium*, Sydney, Australia , 2015, p. 12:1--12:8.

[5] S. Clinchant and F. Perronnin, “Aggregating Continuous Word Embeddings for Information Retrieval,” in *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria, 2013, pp. 100–109.

[6] J. Zhang, S. Liu, M. Li, M. Zhou, and C. Zong, “Towards Machine Translation in Semantic Vector Space,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 14, no. 2, p. 9:1--9:26, Apr. 2015.

[7] S.-Z. Aftabi and A.-M. Chahooki, “Bridging the semantic gap in question classification by categorization rules,” *Tabriz J. Electr. Eng.*, vol. 46, no. 3, pp. 13–24, 2016.

[8] R. Khodaei, M. A. Balafari, and S. N. Razavi, “Effectiveness of Query Expansion based on Clustering of Pseudo-Feedback Documents with K-NN Algorithm,” *Tabriz J. Electr. Eng.*, vol. 46, no. 1, pp.

Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, 2013, pp. 3111–3119.

[24] Q. Le and To. Mikolov, "Distributed Representations of Sentences and Documents," in *The 31st International Conference on Machine Learning*, Beijing, China, 2014, pp. 1188–1196.

[25] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, {EMNLP} 2014, October 25-29, 2014, Doha, Qatar*, 2014, pp. 1532–1543.

[26] Y.-Y. Lee, H. Ke, H.-H. Huang, and H.-H. Chen, "Less is More: Filtering Abnormal Dimensions in GloVe," in *Proceedings of the 25th International Conference Companion on World Wide Web*, Montreal, Canada, 2016, pp. 71–72.

[27] G. Szarvas, C. Biemann, and I. Gurevych, "Supervised All-Words Lexical Substitution using Delexicalized Features," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 2013, pp. 1131–1141.

[28] O. Melamud, O. Levy, and I. Dagan, "A Simple Word Embedding Model for Lexical Substitution," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, Berlin, Germany, 2015, pp. 1–7.

[29] S. Roller and K. Erk, "PIC a Different Word: A Simple Model for Lexical Substitution in Context," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA, 2016, pp. 1121–1126.

[30] X. Wang and C. Zhai, "Mining Term Association Patterns from Search Logs for Effective Query Reformulation," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, Napa Valley, CA, USA, 2008, pp. 479–488.

[31] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating Query Substitutions," in *Proceedings of the 15th International Conference on World Wide Web*, Southampton, UK, 2006, pp. 387–396.

[32] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating Web Spam with Trustrank," in *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, Toronto, Canada, 2004, pp. 576–587.

[33] H. Rubenstein and J. B. Goodenough, "Contextual Correlates of Synonymy," *Commun. ACM*, vol. 8, no. 10, pp. 627–633, Oct. 1965.

[34] J. Camacho-Collados, M. T. Pilehvar, and R. Navigli, "A Framework for the Construction of Monolingual and Cross-lingual Word Similarity Datasets," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China, 2015, pp. 1–7.

[35] D. Inkpen, "A Statistical Model for Near-synonym Choice," *ACM Trans. Speech Lang. Process.*, vol. 4, no. 1, p. 2:1--2:17, Feb. 2007.

[36] X. Xue, V. Dang, and W. B. Croft, "Query Substitution based on N-gram Analysis," CIIR Technical Report, University of Mass. at Amherst, MA, USA, 2009.

143–151, 2016.

[9] D. Yang and D. M. W. Powers, "Measuring Semantic Similarity in the Taxonomy of WordNet," in *Proceedings of the Twenty-eighth Australasian Conference on Computer Science - Volume 38*, 2005, pp. 315–322.

[10] B. B. Klebanov, "Measuring Semantic Relatedness Using People and WordNet," in *Proc. of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, Stroudsburg, PA, USA, 2006, pp. 13–16.

[11] I. Gurevych, "Using the Structure of a Conceptual Network in Computing Semantic Relatedness," in *Proceedings of the Second International Joint Conference on Natural Language Processing*, Berlin, Heidelberg , 2005, pp. 767–778.

[12] T. Costa and J. P. Leal, "Challenges in Computing Semantic Relatedness for Large Semantic Graphs," in *Proceedings of the 18th International Database Engineering & Applications Symposium*, New York, NY, USA, 2014, pp. 376–377.

[13] M. T. Pilehvar and R. Navigli, "From senses to texts: An all-in-one graph-based approach for measuring semantic similarity," *Artif. Intell.*, vol. 228, pp. 95–128, 2015.

[14] T. T. A. Nguyen and S. Conrad, "A Semantic Similarity Measure Between Nouns Based on the Structure of Wordnet," in *Proceedings of International Conference on Information Integration and Web-based Applications and Services*, New York, NY, USA, 2013, pp. 605--609.

[15] P. D. Turney, "Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL," in *Proceedings of the 12th European Conference on Machine Learning*, London, UK, 2001, pp. 491–502.

[16] K. Gulordava and M. Baroni, "A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus," in *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Stroudsburg, PA, USA, 2011, pp. 67–71.

[17] R. L. Cilibrasi and P. M. B. Vitanyi, "The Google Similarity Distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

[18] S. T. Dumais, "Latent semantic analysis," *Annu. Rev. Inf. Sci. Technol.*, vol. 38, no. 1, pp. 188–230, 2004.

[19] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using Latent Semantic Analysis to Improve Access to Textual Information," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 1988, pp. 281–285.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[21] J. Vayrynen and T. Honkela, "Comparison of independent component analysis and singular value decomposition in word context analysis," in *International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, Espoo, Finland, 2005.

[22] Z. Harris, "Distributional structure," *Distrib. Struct.*, vol. 10, no. 2–3, pp. 1456–1162, 1954.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Conference on*

[37] C. Joubarne and D. Inkpen, "Comparison of Semantic Similarity for Different Languages Using the Google N-gram Corpus and Second- Order Co-occurrence Measures," in *Proceedings of the 24th Canadian Conference on Advances in Artificial Intelligence*, St. John's. Canada, 2011, pp. 216–221.

[38] E. Yeh, D. Ramage, C. D. Manning, E. Agirre, and A. Soroa, "WikiWalk: Random Walks on Wikipedia for Semantic Relatedness," in *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, Singapore, 2009, pp. 41–49.

[39] J. Camacho-Collados, M. T. Pilehvar, and R. Navigli, "Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities," *Artif. Intell.*, vol. 240, pp. 36–64, 2016.

زیرنویس‌ها

[1] Substitutability
[2] Lexical context
[3] Semantic similarity
[4] Pointwise Mutual Information (PMI)
[5] Google 1T Ngram
[6] Sinlge Value Decomposition (SVD)
[7] Latent Semantic Analysis (LSA)
[8] Distributed Semantic Model (DSM)
[9] Direct substitutability
[10] Indirect Substitutability
[11] Outlier
[12] Seed nodes
[13] Damping factor
[14] Ngram
[15] Diameter
[16] Sink nodes
[17] Main Function
[18] Characters
[19] Trigram
[20] Quadrigram
[21] Map-Reduce
[22] Hadoop
[23] Small world
[24] Multithread
[25] ε کوچک‌ترین مقدار اعشاری مثبت قابل نمایش در بازه ۰ تا ۱ در هر زبان برنامه‌سازی.