7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

Measurement of FarsNet-based Semantic Word Similarity and Relatedness

Razie Adelkhah*, Mehrnoush Shamsfard

Faculty of Computer Engineering and Science, Shahid Beheshti University
r_adelkhah@sbu.ac.ir, m-shams@sbu.ac.ir

Abstract
Finding similarity and semantic relatedness between words and concepts of a language is very important in natural language processing and can help improve the performance of various systems such as plagiarism detection, summarization, machine translation evaluation, transliteration detection, implication detection and intelligent conversation. Finding semantic similarity and relatedness, depending on the type of meaning representation, can be graph-based or vector-based. In graph-based methods determine the degree of semantic similarity of the two concepts based on the information in the hierarchy, and semantic relatedness is calculated using more information, such as other non-hierarchical relations and glosses or examples for each concept in the wordnet. In this paper, first we explain how the six existing measures of semantic similarity and the three measures of semantic relatedness work on a pair of Persian concepts or words. Besides using these measures, we introduce a new FarsNet-based method and measure semantic similarity and relatedness of Persian words based on all these measures. We also prepare a baseline service to calculate word similarities and test, evaluate or compare similarity measures.

Keywords: Semantic Similarity, Semantic Relatedness, FarsNet

7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

سنجش شباهت و ارتباط معنایی کلمات و مفاهیم فارسی مبتنی بر فارس‌نت

راضیه عادل‌خواه[1*]، مهرنوش شمس‌فرد[2]

[1] آزمایشگاه پردازش زبان طبیعی، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران
r_adelkhah@sbu.ac.ir

[2] آزمایشگاه پردازش زبان طبیعی، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران
m-shams@sbu.ac.ir

چکیده

یافتن شباهت و ارتباط معنایی میان کلمات و مفاهیم یک زبان دارای اهمیت بسیار بالایی در پردازش زبان طبیعی است و می‌تواند به بهبود عملکرد سامانه‌های مختلف مانند کشف تقلب، خلاصه‌سازی، ارزیابی ترجمه ماشینی، تشخیص دگرنویسی، شناسایی استلزام و گفتگوی هوشمند کمک شایانی نماید. یافتن شباهت و ارتباط معنایی بسته به نوع بازنمایی معنا می‌تواند مبتنی بر گراف یا مبتنی بر بردار باشد. در روش‌های مبتنی بر گراف، شباهت معنایی براساس اطلاعات موجود در سلسله‌مراتب، میزان نزدیک بودن دو مفهوم را تعیین می‌نماید و ارتباط معنایی از اطلاعات بیشتری مثل سایر روابط غیر سلسله‌مراتبی و همچنین توضیح و مثال موجود برای هر مفهوم در وردنت‌ها استفاده می‌نماید. در این مقاله پس از توضیح نحوه عملکرد شش معیار شباهت معنایی و سه معیار ارتباط معنایی موجود، برروی زوج مفاهیم و یا زوج واژه‌های زبان فارسی، به معرفی یک روش جدید مبتنی بر فارس‌نت برای این زبان می‌پردازیم. به‌این‌ترتیب ضمن معرفی شیوه جدیدی برای شباهت سنجی کلمات و مفاهیم فارسی مبتنی بر همه روابط فارس‌نت، بستری برای آزمون و ارزیابی و مقایسه روش‌های شباهت سنجی و ارتباط سنجی و همچنین سرویسی برای محاسبه میزان شباهت و ارتباط واژه‌ها یا مفاهیم فراهم آورده‌ایم که در طی مقاله جزئیات آن را بررسی خواهیم کرد.

کلمات کلیدی

شباهت معنایی، ارتباط معنایی، فارس‌نت

۱- مقدمه

شباهت[1] و ارتباط[2] معنایی واژه‌ها از مفاهیم کلیدی در حوزه پردازش زبان طبیعی محسوب می‌گردند. مسئله اصلی، تعیین مقدار عددی میزان نزدیک و یا دور بودن دو واژه و یا مفهوم از یکدیگر است. به عنوان مثال ارتباط معنایی واژه سیب با موز بیشتر از واژه سیب با کتاب است. شباهت و ارتباط معنایی واژه‌ها در کاربردهای بسیاری در حوزه‌های مختلف پردازش زبان طبیعی دارد که از آن جمله می‌توان به ترجمه ماشینی، بازیابی اطلاعات، تشخیص تقلب، تصحیح املا و رفع ابهام معنایی واژه و یادگیری هستان‌شناسی اشاره کرد. معیارها و مجموعه داده‌های متنوع و گسترده‌ای در زبان‌های مختلف برای محاسبه شباهت و ارتباط معنایی واژه‌ها تعریف و تولید شده است. به گفته [5] Taieb ارتباط معنایی را می‌توان میزان مرتبط و نزدیک بودن دو مفهوم در ذهن انسان به یکدیگر دانست و خوب است اشاره شود که تمایز شباهت با ارتباط معنایی در این است که ارتباط معنایی، کلان‌تر و دربرگیرنده شباهت نیز هست و روش‌های مبتنی بر روش‌های توزیعی بیشتر ارتباط معنایی را بررسی

می‌کنند در حالی که روش‌هایی که از هستان‌شناسی‌ها و پایگاه‌های دانش و واژگان استفاده می‌کنند، معمولا شباهت معنایی را محاسبه می‌نمایند.

از اولین کارهای انجام شده در این زمینه می‌توان به مقاله Pedersen [1] اشاره کرد که در آن با استفاده از روابط تعریف شده میان مفاهیم در وردنت [19] یک مجموعه نرم‌افزاری شامل ۶ روش محاسبه شباهت و ۳ روش محاسبه ارتباط معنایی برای زبان انگلیسی ارائه شد.

رویکرد مبتنی بر گراف و استخراج طول مسیر و عمق در وردنت نیز مورد استفاده بوده است [20].

اخیرا Almousa و همکاران [4] برای محاسبه شباهت معنایی در وردنت از روابط غیر سلسله‌مراتبی استفاده کرده‌اند. مدل‌های معنای برداری[3] نیز توسط Quiroz-Mercado و همکاران [6] استفاده شده است که واژه‌ها را به صورت نقاطی در یک فضای با ابعاد بسیار بالا بازنمایی می‌کند. که این روش نیز بر اساس بافتار و با استفاده از گراف دانش ConcepNet شباهت معنایی را محاسبه می‌نماید.

درحالی‌که بیشتر روش‌های محاسبه شباهت مبتنی بر وردنت‌ها تنها از روابط سلسله‌مراتبی شمول (is-a) استفاده می‌کنند. Zhu و همکاران [7] روشی را ارائه کرده‌اند که سایر روابط معنایی را در ترکیب با مفهوم نزدیک‌ترین نسل مشترک مورد استفاده قرار می‌دهد که بدین ترتیب برای واژه‌های قید و صفت نیز می‌تواند شباهت را محاسبه نماید.

در زبان فارسی نیز تاکنون کارهایی در این زمینه انجام گرفته است. رنجبر و همکاران در پروژه مهتاب روشی ترکیبی با استفاده از فارس‌نت و BableNet و Word2Vec ارائه کرده‌اند [3]. برزگر و همکاران، مجموعه داده چندزبانه برای شباهت و ارتباط معنایی برای ۱۱ زبان از جمله زبان فارسی تولید کرده‌اند [8]. پیله‌ور و همکاران از طول مسیر و عمق مفاهیم در سلسله‌مراتب برای محاسبه شباهت استفاده کرده‌اند [20].

در این مقاله میزان شباهت و ارتباط معنایی میان مفاهیم فارس‌نت محاسبه می‌شود. فارس‌نت، بزرگ‌ترین وردنت فارسی است و نسخه سوم آن که مورد استفاده ما در این مقاله است دارای بیش از ۱۰۰ هزار مدخل واژگانی و حدود ۴۰ هزار مجموعه ترادف می‌باشد. برای هر مدخل حداقل یک معنی تعریف شده و هر معنی در یک و فقط یک مجموعه ترادف شرکت می‌کند [18].

برای این منظور، ۶ معیار شباهت و ۳ معیار ارتباط معنایی محاسبه شده در وردنت پرینستون [1] برای فارس‌نت و روابط موجود میان مفاهیم آن، بازپیاده‌سازی شده و مقادیر آن محاسبه شده است. علاوه بر این یک روش نو مبتنی بر استفاده از تمامی روابط معنایی و ویژگی‌های منحصربه‌فرد موجود در فارس‌نت ، ارائه گردیده است. یک از مشکلات بکارگیری روش‌های شباهت سنجی موجود زمان‌گیر بودن آنها به جهت نیاز به پیمایش بخش بزرگی از گراف فارس‌نت است. درحقیقت اهمیت کار انجام شده در این مقاله بیشتر از نوآوری در روش به اجرای همه روش‌های موجود و حتی روش جدید زمان‌بری که از همه روابط موجود در فارس‌نت استفاده می‌کند یک بار برای همه مفاهیم و کلمات فارس‌نت و ذخیره نتایج است تا از این پس محققین بتوانند از نتایج این شباهت سنجی به‌صورت برون خط استفاده نمایند. نتیجه شباهت سنجی ها برروی بیش از ۲۰ هزار زوج اسم–اسم و فعل–فعل فارس‌نت اجرا شده است. این محاسبه بر روی کامپیوترهای آزمایشگاه حدود یکسال بطول انجامیده است و بیش از ۴۲۰ میلیون زوج به ازای هر روش شباهت در پایان اجرا ذخیره می‌شود (۴.۲ میلیارد عدد شباهت پس از تکمیل ۱۰ روش) که می تواند باعث کاهش زمان محاسباتی الگوریتم های دیگر که نیاز به سنجش شباهت مبتنی بر گراف فارس‌نت دارند بشود.

بدین ترتیب سرویس ارائه شده به ازای دو مفهوم و یا دو واژه یک مقدار عددی که بیانگر میزان شباهت و یا ارتباط معنایی آن دو مفهوم است را برای هر روش محاسبه بر میگرداند.

بخش‌های ۲ و ۳ به ترتیب شامل تعریف و نحوه پیاده‌سازی معیارهای شباهت و ارتباط معنایی متناظر با WordNet::Similarity [1] هستند. بخش ۴ به توضیح محاسبه روش ارائه شده برای فارس‌نت می‌پردازد. بخش‌های ۵ و۶ و ۷ به ترتیب شامل ارزیابی و اطلاعات پیکره و ابزار و نتیجه می‌باشد.

۲- معیارهای شباهت معنایی

شباهت معنایی در وردنت‌ها می‌تواند بر اساس طول مسیر میان دو مفهوم و یا مبتنی بر اطلاعات محتوایی همراه آن محاسبه شود.

۲-۱- معیارهای مبتنی بر طول مسیر

Lch -۱-۱-۲

روش معرفی شده توسط Leacock و Chodorow, [9] تعداد یال‌های بین دو گره را در سلسله‌مراتب is-a شمارش می‌کند. سپس عدد به دست آمده را با استفاده از بیشینه عمق موجود در آن سلسله‌مراتب مقیاس می‌نماید. درنهایت میزان شباهت، بر اساس فرمول (۱) منفی لگاریتم عدد به دست آمده خواهد بود.

بیشینه عمق سلسله‌مراتب براساس نسخه مورد استفاده از فارس‌نت ممکن است متغیر باشد ولی این مقدار برای ریشه یکتای سلسله‌مراتب is-a درآخرین نسخه فارس‌نت برابر با ۲۱ است.

$$\text{relatedness} = -\log\frac{L}{2 * D} \qquad (۱)$$

که L همان طول کوتاه‌ترین مسیر در سلسله‌مراتب است. D نیز عمق گره دارای طولانی‌ترین مسیر از ریشه یا همان عمق بیشینه واژگان است.

Wup -۲-۱-۲

این معیار براساس شمارش گره‌هاست [10] و با توجه عمق دو مفهوم در سلسله‌مراتب و همچنین عمق گره LCS[4] آن دو مطابق فرمول (۲) محاسبه می‌شود. LCS دو گره، عمیق‌ترین گره در سلسله‌مراتب است که نیای هر دو گره باشد. ازآنجاکه عمق گره LCS هرگز صفر نیست (عمق ریشه

سلسله‌مراتب یک است) پس هیچ‌گاه شباهت صفر نخواهد شد. و اگر دو مجموعه‌ترادف یکی باشند آنگاه مقدار شباهت یک خواهد شد.

$$\text{relatedness} = \frac{2 * \text{depth(lcs)}}{\text{depth(synset1)} + \text{depth(synset1)}} \quad (۲)$$

Path -۳-۱-۲

این الگوریتم، ارتباط معنایی یا شباهت دو مجموعه‌ترادف را با شمارش تعداد گره‌های موجود در کوتاه‌ترین مسیر میان آن دو در سلسله‌مراتب is-a موجود در فارس‌نت محاسبه می‌کند. طول مسیرها شامل گره پایانی نیز می‌شود. ازآنجاکه مسیر طولانی‌تر نشان‌دهنده ارتباط معنایی کمتر است، عدد خروجی که نشان‌دهنده میزان شباهت است معکوس طول کوتاه‌ترین مسیر بین دو مفهوم است که مطابق فرمول (۳) محاسبه می‌شود. چنانچه دو مفهوم یکی باشند، فاصله آن دو یک خواهد بود پس شباهت نیز یک خواهد شد. همچنین رابطه‌ها بدون جهت در نظر گرفته‌شده‌اند.

$$\text{relatedness} = \frac{1}{\text{distance}} \quad (۳)$$

۲-۲- معیارهای مبتنی بر محتوای اطلاعاتی[۵]

محتوای اطلاعاتی یک معیار مبتنی بر پیکره، برای ارزیابی میزان specificity یک مفهوم است. برای محاسبه اطلاعات محتوایی، دو روش مبتنی بر پیکره[۶] و مبتنی بر رده‌بندی[۷] وجود دارد که روش دوم مورد استفاده قرار گرفته است.

در این روش اطلاعات محتوایی مفهوم c برابر است با نسبت تعداد برگ‌هایی که در سلسله‌مراتب زیر مفهوم c قرار دارند به تعداد مفاهیمی که نیای مفهوم c هستند [11]. درنهایت این مقدار با استفاده از اطلاعات محتوایی ریشه که دارای کمترین اطلاعات محتوایی است نرمال می‌شود و منفی لگاریتم آن با فرمول (۴) محاسبه می‌گردد. اطلاعات محتوایی ریشه برابر با تعداد همه برگ‌ها در سلسله‌مراتب است.

$$\text{IC(c)} = -\log \frac{\frac{|\text{leaves(c)}| + 1}{|\text{subsummer(c)}| + 1}}{\text{maxLeaves} + 1} \quad (۴)$$

Leaves(c): مجموعه مفاهیم موجود در آخرین عمق درخت سلسله‌مراتب (برگ‌ها) است که زیر مفهوم c نیز واقع شده باشند.

Subsummer(c): مجموعه مفاهیمی که در سلسله‌مراتب به عنوان نیا یا اجداد مفهوم c حضور دارند. همچنین این مجموعه شامل خود مفهوم c نیز هست.

maxLeaves: تعداد همه مفاهیمی که در سلسله‌مراتب به عنوان برگ حضور دارند.

res -۱-۲-۲

این روش نیز مبتنی بر محتوای اطلاعاتی دو مفهوم است[12]. این روش شباهت دو مفهوم را طبق فرمول (۵) برابر با محتوای اطلاعاتی گره LCS آن دو مطرح کرده است که مقدار آن بین ۰ و یک خواهد بود.

$$\text{relatedness} = \text{IC(lcs)} \quad (۵)$$

Lin -۲-۲-۲

این معیار از محتوای اطلاعاتی مفاهیم و نظریه شباهت[۸] استفاده می‌کند [13]. در این روش مقدار شباهت برابر است با دو برابر نسبت اطلاعات محتوایی گره LCS دو مفهوم به مجموع اطلاعات محتوایی آن دو که مقدار طبق فرمول (۶) زیر محاسبه می‌شود.

$$\text{relatedness} = \frac{2 * \text{IC(lcs)}}{\text{IC(synset1)} + \text{IC(synset2)}} \quad (۶)$$

IC(lcs) محتوای اطلاعاتی مفهومی است که نقش LCS دو مفهوم موردنظر را داراست.

Jcn -۳-۲-۲

این روش هم با استفاده از محتوای اطلاعاتی دو مفهوم، مطابق فرمول (۷)، شباهت آن دو را محاسبه می‌کند [14].

$$\text{relatedness} = \frac{1}{\text{jcnDistance}} \quad (۷)$$

که jcnDistance با فرمول (۸) به دست می‌آید.

$$\text{jcnDistance} = \text{IC(synset1)} + \text{IC(synset2)} - 2 * \text{IC(lcs)} \quad (۸)$$

۳- معیارهای ارتباط معنایی

Vector -۱-۳

این الگوریتم با استفاده از روشPatwardhan [15] پیاده‌سازی شده است. این معیار، شباهت مفاهیم را با استفاده از بردار هم‌وقوعی مرتبه دوم یا بردار زمینه[۹] برای توضیح[۱۰] مفاهیم محاسبه می‌کند. بردارهای هم‌وقوعی مرتبه دوم با استفاده از پیکره متشکل از جملات توضیح متناظر با مجموعه‌های ترادف موجود در فارس‌نت محاسبه شده‌اند. هر واژه موجود در توضیح دارای یک بردار زمینه متناظر است. هر توضیح نیز دارای بردار متناظری متشکل از میانگین همه بردارهای واژه‌هایش است. میزان رابطه معنایی در این روش با محاسبه کسینوس بردارهای توضیح به دست می‌آید. علاوه بر توضیح خود

مفاهیم، از توضیح مفاهیم دارای رابطه با مفهوم موردنظر نیز برای محاسبه این معیار شباهت استفاده می‌شود. بدین ترتیب با ترکیب توضیح همه مفاهیم دارای رابطه با مفهوم موردنظر، یک توضیح واحد طولانی[11] ساخته می‌شود. درنهایت بردارهای زمینه متناظر با این معنای جدید ساخته شده و شباهت محاسبه می‌شود.

Lesk –۲–۳

این روش از اشتراکات واژه‌ها در توضیح هر مفهوم استفاده می‌کند [16] . به این ترتیب برای هر دو مفهوم موردنظر، اشتراک‌های جمله توضیح برای خود آنها و مفاهیم پدر و فرزند آن دو محاسبه می‌شود. برای محاسبه مقدار اشتراک برای هر دو جمله از روش زیر استفاده می‌شود. ابتدا همه ایست‌واژه‌ها حذف‌شده و تک‌تک واژه‌ها ریشه‌یابی می‌شوند [21]. سپس در هر تکرار، طولانی‌ترین زیرمجموعه مشترک واژه‌ها استخراج می‌گردد. امتیاز هر اشتراک در هر مرحله، مربع طول زیرمجموعه مشترک در آن مرحله است. و درنهایت امتیاز همه مراحل باهم جمع خواهد شد. خروجی نهایی، مجموع امتیاز حاصل از اشتراک معانی است.

Hso –۳–۳

این روش از کوتاه‌ترین مسیر بین دو مفهوم استفاده می‌کند. اما مسیرها فقط از روابط سلسله‌مراتبی تشکیل نشده‌اند. همه انواع رابطه‌ها مجاز شمرده می‌شوند. به‌طورخلاصه در این روش به هریک از انواع رابطه‌های موجود در فارس‌نت یک جهت داده می‌شود. سپس کوتاه‌ترین مسیر بین دو مفهوم که دارای کمترین تغییر جهت باشد، به‌عنوان بهترین مسیر تعیین می‌گردد [17].

٤– روش پیشنهادی

در این روش برای محاسبه شباهت‌ها از همه روابط معنایی موجود در فارس‌نت تا چندین سطح استفاده‌شده است. انواع روابط معنایی تحت پوشش فارس‌نت عبارتند از:

- شمول و زیر شمول
- جزء واژگی و کل واژگی
- علیت
- استلزام منطقی
- رابطه اشتقاقی
- رابطه میان صفت و اسم ویژگی
- رابطه میان اسم و صفت برجسته یا بالقوه
- رابطه میان فعل و آرگومان‌های آن
- رابطه بی‌نام
- نگاشت میان زبانی

همچنین در فارس‌نت اطلاعات آرگومان‌های نحوی و معنایی (قاب فعل) افعال ساده زبان نیز درج شده است [18] .

این روش دو واژه را دریافت نموده و ابتدا همه مجموعه ترادف‌هایی را که شامل آن دو واژه است استخراج می‌نماید. یک معیار فاصله میان هر دو واژه ورودی x و y به آنها نسبت داده می‌شود و محاسبه شباهت با استفاده از این فاصله و بر اساس فرمول (۹) محاسبه می‌شود [2] .

$$\text{similarity}(x,y) = \begin{cases} B^{-aD(x,y)}, D(x,y) < 0 \\ 0, \text{otherwisse} \end{cases} \qquad (۹)$$

D(x,y): فاصله نسبت داده شده برای واژه ورودی x و y است که بخش اصلی روش ارائه‌شده محاسبه مقدار این فاصله به ازای هر دو واژه دلخواه ورودی است. مقدار پارامترهای a و B بر اساس مرجع [3]، 0.25 و 1 در نظرگرفته شده است.

٤–۱– محاسبه فاصله دو واژه

همان‌طور که اشاره شد ابتدا فاصله دو واژه از هم محاسبه می‌گردد $D(x,y)$. برای این کار واژه‌ها به عنوان ورودی دریافت می‌شوند. سپس تعداد اشتراک‌های میان مجموعه ترادف‌هایی که دو واژه به آنها تعلق دارند تعیین می‌گردد.

چنانچه مجموعه مشترکی میان مجموعه‌های ترادف دو واژه وجود داشته باشد،آنگاه با توجه به تعریف مجموعه ترادف، دو واژه دارای یک یا بیشتر معنای مشترک هستند و بر اساس تعداد این اشتراک‌ها و تعداد همه مجموعه‌های ترادف هر واژه، مقدار فاصله مطابق زیر تعیین می‌گردد.

- اگر هر دو واژه دارای یک مجموعه ترادف باشند: D(x,y)= 0
- اگر فقط یکی از واژه‌ها دارای یک مجموعه ترادف باشد: D(x,y)= 0
- اگر هر دو دارای بیش از یک مجموعه ترادف باشند: D(x,y)= 0.5

اگر مجموعه‌های ترادف دو واژه با هم اشتراکی نداشته باشند، فرایند محاسبه فاصله مطابق زیر و به ترتیب کامل می‌گردد.

چنانچه رابطه سلسله‌مراتبی شمول/زیرشمول میان مجموعه‌های ترادف وجود داشته باشد از فرمول (۱۰) استفاده می‌شود.

$$D(x,y) = 1 + (n-1) * 0.5 \qquad (۱۰)$$

n تعداد سطوح فاصله موجود میان مجموعه‌های ترادف است که برای رابطه مستقیم برابر با یک می‌باشد.

چنانچه مجموعه‌های دارای رابطه سلسله‌مراتبی شمول مستقیم با مجموعه‌های ترادف موردنظر، با هم اشتراک داشته باشند:

- اگر هر دو دارای فقط یک والد باشند: D(x,y)= 1

- اگر یکی از آنها دارای بیش از یک والد باشد: D(x,y)= 1.5
- اگر هر دو دارای بیش از یک والد باشند: D(x,y)= 2

چنانچه مجموعه‌های دارای رابطه سلسله‌مراتبی شمول با فاصله بیش از یک سطح (تا حداکثر پنج سطح) با مجموعه‌های ترادف موردنظر با هم اشتراک داشته باشند از فرمول (۱۱) برای محاسبه شباهت استفاده می‌شود.

$$D(x,y) = 2 + n * 0.5 \qquad (۱۱)$$

n تعداد سطوح فاصله موجود میان مجموعه‌های ترادف دارای رابطه سلسله‌مراتبی شمول غیرمستقیم با مجموعه‌های ترادف موردنظر است.

چنانچه مجموعه‌های دارای رابطه سلسله‌مراتبی شمول با فاصله‌های مختلف (از یک تا پنج سطح) با مجموعه‌های ترادف موردنظر با هم اشتراک داشته باشند:

- اگر فاصله بیشتر برابر با دو یا سه سطح باشد: D(x,y)= 3
- اگر فاصله بیشتر برابر با چهار یا پنج سطح باشد: D(x,y)= 4

چنانچه مجموعه‌های دارای رابطه سلسله‌مراتبی زیرشمول مستقیم با مجموعه‌های ترادف موردنظر، با هم اشتراک داشته باشند: D(x,y)= 1.5

چنانچه مجموعه‌های دارای رابطه سلسله‌مراتبی زیرشمول با فاصله بیش از یک سطح (تا حداکثر پنج سطح) با مجموعه‌های ترادف موردنظر با هم اشتراک داشته باشند، از فرمول (۱۲) برای محاسبه شباهت استفاده می‌شود.

$$D(x,y) = 2.5 + n * 0.5 \qquad (۱۲)$$

n تعداد سطوح فاصله موجود میان مجموعه‌های ترادف دارای رابطه سلسله‌مراتبی زیرشمول غیرمستقیم با مجموعه‌های ترادف موردنظر است.
چنانچه مجموعه‌های دارای رابطه سلسله‌مراتبی زیرشمول با فاصله‌های مختلف (از یک تا پنج سطح) با مجموعه‌های ترادف موردنظر با هم اشتراک داشته باشند:

- اگر فاصله بیشتر برابر با دو یا سه سطح باشد: D(x,y)= 3
- اگر فاصله بیشتر برابر با چهار یا پنج سطح باشد: D(x,y)= 4

چنانچه هر یک از رابطه‌های غیر شمول ممکن میان مجموعه‌های ترادف[۱۲] بین آن‌ها وجود داشته باشد:

- اگر رابطه Related-to بین آن دو وجود داشته باشد: D(x,y)= 1.5
- اگر رابطه Domain بین آن دو وجود داشته باشد: D(x,y)= 3
- اگر سایر رابطه‌ها بین آن دو وجود داشته باشد: D(x,y)= 2

چنانچه هر یک از رابطه‌های ممکن میان معانی موجود برای هر واژه بین آنها وجود داشت:

- اگر Co_Occurrence یا Antonym یا Derivationally_related_form بود: D(x,y)= 2
- اگر سایر رابطه‌ها[۱۳] بین آن دو وجود داشته باشد: D(x,y)= 3

چنانچه میان مجموعه‌های دارای رابطه‌های غیرشمول با مجموعه‌های ترادف موردنظر با فاصله یک سطح، رابطه‌ای وجود داشته باشد: D(x,y)= 3

چنانچه میان مجموعه‌های دارای رابطه‌های غیرشمول با مجموعه‌های ترادف موردنظر با فاصله دو سطح، رابطه ای وجود داشته باشد: D(x,y)= 4

مقادیر برای هر یک از شرایط ذکر شده با سعی و خطا بر اساس بهترین نتیجه حاصل از داده مورد ارزیابی به دست آمده است.

۵- ارزیابی

ارزیابی بر روی داده‌های SEMEVAL 2017 انجام شده است که شامل ۵۰۰ جفت واژه می‌باشد که مقدار عددی شباهت معنایی آنها تعیین شده است. برای مقایسه و ارزیابی از معیار ضریب همبستگی که از معیارهای مورد استفاده در تعیین همبستگی دو متغیر است استفاده شده و نتایج در جدول (۱) قابل مشاهده است.
ابتدا شباهت هر زوج واژه با استفاده از روش‌های معرفی شده محاسبه گردید و یک بردار به طول ۵۰۰ ساخته شده است. این همبستگی میان بردار شباهت معنایی Gold در مجموعه داده SEMEVAL 2017 به طول ۵۰۰ (به ازای ۵۰۰ زوج واژه در داده مورد ارزیابی) با بردار شباهت معنایی حاصل از هر روش به صورت جداگانه محاسبه شده است تا شدت و نوع رابطه (مستقیم یا معکوس) نتایج حاصل از هر یک از روش‌ها با مقادیر gold را نشان دهد.
ازآنجاکه مقادیر شباهت در پیکره برای مفاهیم محاسبه شده‌اند و نه واژه‌ها، به ازای هر واژه، شباهت میان همه مجموعه ترادف‌هایی که شامل آن واژه‌ها هستند به صورت دوبه دو محاسبه شده (ضرب کارتزین) و مقدار بیشینه به عنوان شباهت دو واژه گزارش شده است.
در مقایسه نتایج با روش مهتاب [3] به عنوان برنده چالش SEMEVAL 2017 روی شباهت سنجی فارسی، باتوجه به اینکه در آن مقاله از ترکیب ۳ روش مبتنی بر رابطه شمول مستقیم در فارس‌نت و bableNet و Word2Vec استفاده شده است، برای مقایسه دقیق‌تر دقت روش پیشنهادی در این مقاله که فقط مبتنی بر فارس‌نت است، مقدار عددی نتیجه حاصل از اجرای فقط بخش مبتنی بر فارس‌نت از آن [3] آورده شده است.

جدول (۱) : نتایج ارزیابی هر یک از معیارهای محاسبه شباهت

measure	Pearson correlation	Spearman correlation	Average correlation
LCH	0.625	0.627	0.626
Wup	0.515	0.546	0.5305
Path	0.591	0.627	0.609
Res	0.585	0.583	0.584
Lin	0.596	0.586	0.591
JCN	0.419	0.572	0.4955
Vector	0.291	0.348	0.3375
Lesk	0.199	0.595	0.397
HSO	0.488	0.647	0.5675
Mahtab (farsnet)	0.371	0.518	0.4445
Proposed method	0.643	0.692	0.6675

٦- اطلاعات پیکره

پس از پیاده‌سازی، معیارهای شباهت و ارتباط معنایی میان همه اسم‌ها با هم و همه فعل‌ها با یکدیگر در فارس‌نت به‌صورت زوج‌های دوتایی محاسبه و در قالب یک پیکره ذخیره شده‌اند. (زوج‌های فعل- فعل و اسم- اسم). که این تعداد بیش از ۴۲۰ میلیون زوج می‌باشد. مقادیر شباهت تا ۴ رقم اعشار نگهداری شده‌اند. همچنین برای سایر روش‌ها امکان دسترسی و استفاده از طریق سرویس تحت وب در نظر گرفته شده است.

۷- جمع بندی

آنچه در این پژوهش مورد توجه بود ارائه مجموعه‌ای برای محاسبه شباهت و ارتباط معنایی واژه‌های فارسی بر اساس روش‌های مبتنی بر دانش و با استفاده از فارس‌نت بوده است. با توجه به اینکه فرایند محاسبه شباهت برای همه واژه‌های زبان فارسی با توجه به ابعاد بسیار بزرگ فارس‌نت بسیار زمان‌بر است، یکی از اهداف اساسی، ایجاد خدمت پایه شباهت سنجی و ایجاد امکان استفاده سهل و سریع از آن در روش‌های دیگر که نیاز به شباهت سنجی واژه‌های زبان فارسی دارند می‌باشد. در راستای این هدف ۹ روش معمول و مورد استفاده در وردنت‌ها و یک روش نو معرفی و پیاده‌سازی گردید. ارزیابی روش ارائه شده نسبت به سایرین توانست امتیاز بالاتری به دست آورد اما برای کارهای آتی استفاده از روش‌های جدید مبتنی بر پیکره و بر اساس بافتار و ترکیب نتایج این دو می‌تواند در بهبود نتایج بسیار موثر باشد.

مراجع

[1] Pedersen, T., Siddharth, P. and Jason, M., *"WordNet:: Similarity-Measuring the Relatedness of Concepts."* , AAAI, Vol. 4, 2004.

[2] Rychalska, B., et al,. *"Samsung poland nlp team at semeval-2016 task 1: Necessity for diversity; combining recursive autoencoders, wordnet and ensemble methods to measure semantic similarity."*, Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016.

[3] Ranjbar, N., Mashhadirajab, F., and Shamsfard, M., *"Mahtab at semeval-2017 task 2: Combination of corpus-based and knowledge-based methods to measure semantic word similarity."* Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017.

[4] AlMousa, M., Benlamri, R., and Khoury R., *"Exploiting non-taxonomic relations for measuring semantic similarity and relatedness in WordNet."*, Knowledge-Based Systems,Vol. 212, 2021.

[5] Taieb, M. A. H., Zesch, T., and Aouicha M., *"A survey of semantic relatedness evaluation datasets and procedures."*, Artificial Intelligence Review, Vol. 53, No. 6, pp. 4407-4448, 2020.

[6] Quiroz-Mercado, Isaias, J., Barrón-Fernández, R., and Ramírez-Salinas, M., *"Semantic Similarity Estimation Using Vector Symbolic Architectures."*, IEEE Access, Vol. 8, pp.109120-109132, 2020.

[7] Zhu, X., et al., *"Measuring similarity and relatedness using multiple semantic relations in WordNet."*, Knowledge and Information Systems, 2019.

[8] Barzegar, S., et al., *"SemR-11: A multi-lingual gold-standard for semantic similarity and relatedness for eleven languages."*, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[9] Leacock, C., and Chodorow, M., *"Combining local context and WordNet similarity for word sense identification."* ,WordNet: An electronic lexical database, Vol. 49, No. 2, pp. 265-283, 1998.

[10] Wu, Z., and Palmer, M., *"Verb semantics and lexical selection*.", arXiv preprint cmp-lg/9406033,1994.

[11] Sánchez, D., Batet, M., and Isern, D., *"Ontology-based information content computation."*, Knowledge-based systems, Vol. 24, pp. 297-303, 2011.

[12] Resnik, P., *"Using information content to evaluate semantic similarity in a taxonomy."* ,arXiv preprint cmp-lg/9511007 ,1995.

[13] Lin, D,. *"An information-theoretic definition of similarity."*, Icml, Vol. 98, 1998.

[14] Jiang, J. J., and Conrath, D. W. , *"Semantic similarity based on corpus statistics and lexical taxonomy."* ,arXiv preprint cmp-lg/9709008 ,1997.

[15] Pedersen, T., Siddharth, P , *Incorporating dictionary and corpus information into a context vector measure of semantic relatedness*, Ph.D. Thesis, University of Minnesota, Duluth, 2003.

[16] Banerjee, S., and Pedersen, T., *"Extended gloss overlaps as a measure of semantic relatedness."*, Ijcai, Vol. 3, 2003.

[17] Hirst, G., and St-Onge, D., *"Lexical chains as representations of context for the detection and correction of malapropisms."* ,WordNet: An electronic lexical database ,Vol. 305, pp. 305-332, 1998.

[18] Shamsfard, M., et al., *"Semi automatic development of farsnet; the persian wordnet."* ,Proceedings of 5th global WordNet conference, Mumbai, India. Vol. 29, 2010.

[19] Miller, G. A., *"WordNet: a lexical database for English."* ,Communications of the ACM, Vol.38, No. 11, pp. 39-41, 1995.
[20] Pilehvar, M. T., and Navigli, R., *"From senses to texts: An all-in-one graph-based approach for measuring semantic similarity."* ,Artificial Intelligence, Vol. 228, pp. 95-128, 2015.
[21] Shamsfard, M., Jafari, H. S., and Ilbeygi, M., *"STeP-1: A Set of Fundamental Tools for Persian Text Processing*.", LREC. 2010.

زیرنویس‌ها

[1] Similarity
[2] Relatedness
[3] Vector Semantic Models
[4] Least Common Subsummer
[5] Information Content
[6] Corpus-based
[7] Taxonomy-based
[8] Similarity Theorem
[9] Context Vector
[10] Gloss
[11] Super Gloss
[12] Antonym, Part holonym, Part meronym, Member holonym, Member holonym, Attribute, is-Attribute-of, Portion meronym, Portion holonym, Salient defining feature, Has-Salient defining feature, Substance holonym, Substance meronym, Entailment, Is-Entailed-by, Is-Patient-of, Patient, Agent, Is-Agent-of, Cause, Is-Caused-by, Is-Location-of, Location, Is-Instrument-of, Instrument, Unit, Has-Unit, Related -to, Domain, Is-Domain-of
[13] Non_Verbal_Part , Refer_to, Is_Verbal_Part_of , Is_Referred_by , Is_Non_Verbal_Part_of