

## لینک های مفید



عضویت  
در خبرنامه



کارگاه های  
آموزشی



سرویس  
ترجمه تخصصی  
STRS



فیلم های  
آموزشی



بلاگ  
مرکز اطلاعات علمی



سرویس های  
ویژه



# 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

## سنجش شباهت و ارتباط معنایی کلمات و مفاهیم فارسی مبتنی بر فارس نت

راضیه عادل خواه<sup>۱\*</sup>، مهرنوش شمس فرد<sup>۲</sup>

<sup>۱</sup> آزمایشگاه پردازش زبان طبیعی، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران  
r\_adelkhah@sbu.ac.ir

<sup>۲</sup> 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، بر روی زوج مفاهیم و یا زوج واژه‌های زبان فارسی، به معرفی یک روش جدید مبتنی بر فارس نت برای این زبان می‌پردازیم. به این ترتیب ضمن معرفی شیوه جدیدی برای شباهت سنجی کلمات و مفاهیم فارسی مبتنی بر همه روابط فارس نت، 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، فارس نت

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

### ۱- مقدمه

شباهت<sup>۱</sup> و ارتباط<sup>۲</sup> 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،

بتوانند از نتایج این شباهت سنجی به صورت برون خط استفاده نمایند. نتیجه شباهت سنجی ها بر روی 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 می شود (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.

بخش های ۲ و ۳ به ترتیب شامل تعریف و نحوه پیاده سازی معیارهای شباهت و ارتباط معنایی متناظر با WordNet::Similarity [1] هستند. بخش ۴ به توضیح محاسبه روش ارائه شده برای فارسنت می پردازد. بخش های ۵ و ۶ و ۷ به ترتیب شامل ارزیابی و اطلاعات پیکره و ابزار و نتیجه می باشد.

## ۲- معیارهای شباهت معنایی

شباهت معنایی در وردنت ها می تواند بر اساس طول مسیر میان دو مفهوم و یا مبتنی بر اطلاعات محتوایی همراه آن محاسبه شود.

### ۲-۱- معیارهای مبتنی بر طول مسیر

#### ۲-۱-۱- Lch

روش معرفی شده توسط Leacock و Chodorow [9] تعداد یال های بین دو گره را در سلسله مراتب is-a شمارش می کند. سپس عدد به دست آمده را با استفاده از بیشینه عمق موجود در آن سلسله مراتب مقیاس می نماید. در نهایت میزان شباهت، بر اساس فرمول (۱) منفی لگاریتم عدد به دست آمده خواهد بود.

بیشینه عمق سلسله مراتب بر اساس نسخه مورد استفاده از فارسنت ممکن است متغیر باشد ولی این مقدار برای ریشه یکنای سلسله مراتب is-a در آخرین نسخه فارسنت برابر با ۲۱ است.

$$\text{relatedness} = -\log \frac{L}{2 * D} \quad (1)$$

که L همان طول کوتاه ترین مسیر در سلسله مراتب است. D نیز عمق گره دارای طولانی ترین مسیر از ریشه یا همان عمق بیشینه واژگان است.

#### ۲-۱-۲- Wup

این معیار بر اساس شمارش گره هاست [10] و با توجه عمق دو مفهوم در سلسله مراتب و همچنین عمق گره  $LCS^*$  آن دو مطابق فرمول (۲) محاسبه می شود. LCS دو گره، عمیق ترین گره در سلسله مراتب است که نیای هر دو گره باشد. از آنجا که عمق گره LCS هرگز صفر نیست (عمق ریشه

می کنند در حالی که روش هایی که از هستان شناسی ها و پایگاه های دانش و واژگان استفاده می کنند، معمولاً شباهت معنایی را محاسبه می نمایند.

از اولین کارهای انجام شده در این زمینه می توان به مقاله [1] Pedersen اشاره کرد که در آن با استفاده از روابط تعریف شده میان مفاهیم در وردنت [19] یک مجموعه نرم افزاری 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].

اخیراً Almousa و همکاران [4] برای محاسبه شباهت معنایی در وردنت از روابط غیر سلسله مراتبی استفاده کرده اند. مدل های معنای برداری نیز توسط Quiroz-Mercado و همکاران [6] استفاده شده است که واژه ها را به صورت نقاطی در یک فضای با ابعاد بسیار بالا بازنمایی می کند. که این روش نیز بر اساس بافتار و با استفاده از گراف دانش ConcepNet شباهت معنایی را محاسبه می نماید.

در حالی که بیشتر روش های محاسبه شباهت مبتنی بر وردنت ها تنها از روابط سلسله مراتبی شمول (is-a) استفاده می کنند. Zhu و همکاران [7] روشی را ارائه کرده اند که سایر روابط معنایی را در ترکیب با مفهوم نزدیک ترین نسل مشترک مورد استفاده قرار می دهد که بدین ترتیب برای واژه های قید و صفت نیز می تواند 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 کرده اند [3]. برزگر و همکاران، مجموعه داده چندزبانه برای شباهت و ارتباط معنایی برای ۱۱ زبان از جمله زبان فارسی تولید کرده اند [8]. پلهور و همکاران از طول مسیر و عمق مفاهیم در سلسله مراتب برای محاسبه شباهت استفاده کرده اند [20].

در این مقاله میزان شباهت و ارتباط معنایی میان مفاهیم فارسنت محاسبه می شود. فارسنت، بزرگ ترین 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 مترادف می باشد. برای هر مدخل حداقل یک معنی تعریف شده و هر معنی در یک و فقط یک مجموعه مترادف شرکت می کند [18].

برای این منظور، ۶ معیار شباهت و ۳ معیار ارتباط معنایی محاسبه شده در وردنت پربنستون [1] برای فارسنت و روابط موجود میان مفاهیم آن، بازپیاده سازی شده و مقادیر آن محاسبه شده است. علاوه بر این یک روش نو مبتنی بر استفاده از تمامی روابط معنایی و ویژگی های منحصر به فرد موجود در فارسنت، ارائه گردیده است. یک از مشکلات بکارگیری روش های شباهت سنجی موجود زمان گیر بودن آنها به جهت نیاز به پیمایش بخش بزرگی از گراف فارسنت است. در حقیقت اهمیت کار انجام شده در این مقاله بیشتر از نوآوری در روش به اجرای همه روش های موجود و حتی روش جدید زمان بری که از همه روابط موجود در فارسنت استفاده می کند یک بار برای همه مفاهیم و کلمات فارسنت و ذخیره نتایج است تا از این پس محققین

### res-1-2-2

این روش نیز مبتنی بر محتوای اطلاعاتی دو مفهوم است [12]. این روش شباهت دو مفهوم را طبق فرمول (5) برابر با محتوای اطلاعاتی گره LCS آن دو مطرح کرده است که مقدار آن بین 0 و یک خواهد بود.

$$\text{relatedness} = \text{IC}(lcs) \quad (5)$$

### Lin-2-2-2

این معیار از محتوای اطلاعاتی مفاهیم و نظریه شباهت<sup>۱۱</sup> 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 (6) زیر محاسبه می‌شود.

$$\text{relatedness} = \frac{2 * \text{IC}(lcs)}{\text{IC}(\text{synset1}) + \text{IC}(\text{synset2})} \quad (6)$$

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

### Jcn-3-2-2

این روش هم با استفاده از محتوای اطلاعاتی دو مفهوم، مطابق فرمول (7)، شباهت آن دو را محاسبه می‌کند [14].

$$\text{relatedness} = \frac{1}{\text{jcnDistance}} \quad (7)$$

که jcnDistance با فرمول (8) به دست می‌آید.

$$\text{jcnDistance} = \text{IC}(\text{synset1}) + \text{IC}(\text{synset2}) - 2 * \text{IC}(lcs) \quad (8)$$

## 3- معیارهای ارتباط معنایی

### Vector-1-3

این الگوریتم با استفاده از روش Patwardhan [15] پیاده‌سازی شده است. این معیار، شباهت مفاهیم را با استفاده از بردار هم‌وقوعی مرتبه دوم یا بردار زمینه<sup>۱۲</sup> برای توضیح<sup>۱۳</sup> مفاهیم محاسبه می‌کند. بردارهای هم‌وقوعی مرتبه دوم با استفاده از پیکره متشکل از جملات توضیح متناظر با مجموعه‌های ترادف موجود در فارسی 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

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

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

### Path-3-1-2

این الگوریتم، ارتباط معنایی یا شباهت دو مجموعه ترادف را با شمارش تعداد گره‌های موجود در کوتاه‌ترین مسیر میان آن دو در سلسله‌مراتب is-a موجود در فارسی محاسبه می‌کند. طول مسیرها شامل گره پایانی نیز می‌شود. از آنجاکه 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 (3) 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 گرفته شده‌اند.

$$\text{relatedness} = \frac{1}{\text{distance}} \quad (3)$$

### 2-2- معیارهای مبتنی بر محتوای اطلاعاتی<sup>۱۴</sup>

محتوای اطلاعاتی یک معیار مبتنی بر پیکره، برای ارزیابی میزان specificity یک مفهوم است. برای محاسبه اطلاعات محتوایی، دو روش مبتنی بر پیکره<sup>۱۵</sup> و مبتنی بر رده‌بندی<sup>۱۶</sup> 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]. در نهایت این مقدار با استفاده از اطلاعات محتوایی ریشه که دارای کمترین اطلاعات محتوایی است نرمال می‌شود و منفی لگاریتم آن با فرمول (4) محاسبه می‌گردد. اطلاعات محتوایی ریشه برابر با تعداد همه برگ‌ها در سلسله‌مراتب است.

$$\text{IC}(c) = -\log \frac{|\text{leaves}(c)| + 1}{|\text{subsumer}(c)| + 1} \frac{1}{\text{maxLeaves} + 1} \quad (4)$$

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

Subsumer(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.

همچنین در فارسی نت 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].

$$\text{similarity}(x, y) = \begin{cases} B^{-aD(x,y)}, D(x, y) < 0 \\ 0, \text{otherwise} \end{cases} \quad (9)$$

$D(x, y)$ : فاصله نسبت داده شده برای واژه ورودی  $x$  و  $y$  است که بخش اصلی روش ارائه شده 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 و نظر گرفته شده است.

#### ۴-۱- 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.

$$D(x, y) = 1 + (n - 1) * 0.5 \quad (10)$$

$n$  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:  $D(x, y) = 1$

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 ساخته می‌شود. در نهایت بردارهای زمینه متناظر با این معنای جدید ساخته شده و شباهت محاسبه می‌شود.

#### ۳-۲- Lesk

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 حذف شده 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 در نهایت امتیاز همه مراحل با هم جمع خواهد شد. خروجی نهایی، مجموع امتیاز حاصل از اشتراک معانی است.

#### ۳-۳- Hso

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

#### ۴- روش پیشنهادی

در این روش برای محاسبه شباهت‌ها از همه روابط معنایی موجود در فارسی نت تا چندین سطح استفاده شده است. انواع روابط معنایی تحت پوشش فارسی نت 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

- 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<sup>۱۱</sup> بین آن دو وجود داشته باشد:  $D(x,y)=3$
- چنانچه میان مجموعه‌های دارای رابطه‌های غیرشامل با مجموعه‌های مترادف موردنظر با فاصله یک سطح، رابطه‌ای وجود داشته باشد:  $D(x,y)=3$
- چنانچه میان مجموعه‌های دارای رابطه‌های غیرشامل با مجموعه‌های مترادف موردنظر با فاصله دو سطح، رابطه‌ای 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 2017 SEMEVAL 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 2017 SEMEVAL 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.

از آنجاکه مقادیر شباهت در پیکره برای مفاهیم محاسبه شده‌اند و نه واژه‌ها، به ازای هر واژه، شباهت میان همه مجموعه‌ترادف‌هایی که شامل آن واژه‌ها هستند به صورت دوجه دو محاسبه شده (ضرب کارترین) 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 شامل مستقیم در فارسی‌نت و bableNet و Word2Vec استفاده شده است، برای مقایسه دقیق‌تر دقت روش پیشنهادی در این مقاله که فقط مبتنی بر فارسی‌نت است، مقدار عددی نتیجه حاصل از اجرای فقط بخش مبتنی بر فارسی‌نت از آن [3] آورده شده است.

- 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 شامل با فاصله بیش از یک سطح (تا حداکثر پنج سطح) با مجموعه‌های مترادف موردنظر با هم اشتراک داشته باشند از فرمول (۱۱) برای محاسبه شباهت استفاده می‌شود.
- $$D(x,y) = 2 + n * 0.5 \quad (11)$$
- n تعداد سطوح فاصله موجود میان مجموعه‌های مترادف دارای رابطه سلسله‌مراتبی شامل غیرمستقیم با مجموعه‌های مترادف موردنظر است.
- چنانچه مجموعه‌های دارای رابطه سلسله‌مراتبی شامل با فاصله‌های مختلف (از یک تا پنج سطح) با مجموعه‌های مترادف 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 زیرشامل مستقیم با مجموعه‌های مترادف موردنظر، با هم اشتراک داشته باشند:  $D(x,y)=1.5$
- چنانچه مجموعه‌های دارای رابطه سلسله‌مراتبی زیرشامل با فاصله بیش از یک سطح (تا حداکثر پنج سطح) با مجموعه‌های مترادف موردنظر با هم اشتراک داشته باشند، از فرمول (۱۲) برای محاسبه شباهت استفاده می‌شود.

$$D(x,y) = 2.5 + n * 0.5 \quad (12)$$

n تعداد سطوح فاصله موجود میان مجموعه‌های مترادف دارای رابطه سلسله‌مراتبی زیرشامل غیرمستقیم با مجموعه‌های مترادف موردنظر است.

چنانچه مجموعه‌های دارای رابطه سلسله‌مراتبی زیرشامل با فاصله‌های مختلف (از یک تا پنج سطح) با مجموعه‌های مترادف 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 شامل ممکن میان مجموعه‌های مترادف<sup>۱۲</sup> 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:  $D(x,y)=3$

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.

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

## ۶- اطلاعات پیکره

پس از پیاده‌سازی، معیارهای شباهت و ارتباط معنایی میان همه اسم‌ها با هم و همه فعل‌ها با یکدیگر در فارسی به صورت 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 فارسی بوده است. با توجه به اینکه فرایند محاسبه شباهت برای همه واژه‌های زبان فارسی با توجه به ابعاد بسیار بزرگ فارسی 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



- [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.

## زیر نویس ها

---

<sup>1</sup> Similarity

<sup>2</sup> Relatedness

<sup>3</sup> Vector Semantic Models

<sup>4</sup> Least Common Subsummer

<sup>5</sup> Information Content

<sup>6</sup> Corpus-based

<sup>7</sup> Taxonomy-based

<sup>8</sup> Similarity Theorem

<sup>9</sup> Context Vector

<sup>10</sup> Gloss

<sup>11</sup> Super Gloss

<sup>12</sup> 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

<sup>13</sup>Non\_Verbal\_Part , Refer\_to, Is\_Verbal\_Part\_of , Is\_Referred\_by , Is\_Non\_Verbal\_Part\_of

## لینک های مفید



عضویت  
در خبرنامه



کارگاه های  
آموزشی



سرویس  
ترجمه تخصصی  
STRS



فیلم های  
آموزشی



بلاگ  
مرکز اطلاعات علمی



سرویس های  
ویژه