Identifying Persian Words' Senses Automatically by Utilizing the Word Embedding Method

Masood Ghayoomi
PhD in Computational Linguistics; Assistant Professor; Institute for Humanities and Cultural Studies Email: M.Ghayoomi@ihcs.ac.ir

Iranian Journal of Information Processing and Management

Received: 28, Dec. 2018 **Accepted: 29, Jul. 2019**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 35 | No. 1 | pp. 25-50
Autumn 2019
https:/doi.org/10.35050/JIPM010.2019.001

Abstract: A word is the smallest unit in a language that has 'form' and 'meaning'. The word might have more than one meaning in which its exact meaning is determined according to the context it is appeared. Collecting all words' senses manually is a tedious and time consuming task. Moreover, it is possible that the words' meanings change over time such that the meaning of an existing word will become unusable or a new meaning will be added to the word. Computational methods is one of the approaches used for identifying words' senses with respect to the linguistic contexts.

In this paper, we put an effort to propose an algorithm to identify senses of Persian words automatically without a human supervision. To reach this goal, we utilize the word embedding method in a vector space model. To build words' vectors, we use an algorithm based on the neural network approach to gather the context information of the words in the vectors. In the proposed model of this research, the divisive clustering algorithm as one of hierarchical clustering algorithms fits with the requirements of our research question. In the proposed model, two modes, namely the Sentence-based and the Context-based, are introduced to identify words' senses. In the Sentence-based mode, all of the words in a sentence that contain the target word are involved to build the sentence vector; while in the Context-based mode, only a limited number of surrounding words of the target word is involved to build the sentence vector. Two evaluation metrics, namely internal and external, are required to evaluate the performance of the clustering algorithm. The silhouette score for each cluster is computed as the internal evaluation metric for both modes of the proposed model. The external evaluation requires a gold standard data for which a data set containing 20 ambiguous words and 100 sentences for each target word is developed.

According to the obtained results of the internal evaluation, the Sentence-based mode has higher density of clusters than the Context-based mode, and the difference between them is statistically significant.

According to the V- and F-measure evaluation metrics in the external evaluation, the Context-based mode has obtained higher performance against the baselines with statistically significant difference.

Keywords: Word Embedding, Clustering, Unsupervised Machine Learning, Vector Space, Natural Language Processing, Word sense representation, Persian

تعیین خودکار معانی واژه‌های فارسی با استفاده از تعبیهٔ معنایی واژه

مسعود قیومی

دکتری زبان‌شناسی رایانشی؛ استادیار؛
پژوهشگاه علوم انسانی و مطالعات فرهنگی؛
M.Ghayoomi@ihcs.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۷/۱۰/۰۷ **پذیرش:** ۱۳۹۸/۰۵/۰۷ مقاله برای اصلاح به مدت ۱۸ روز نزد پدیدآور بوده است.

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۵ | شمارهٔ ۱ | ص‌ص ۲۵-۵۰
پاییز ۱۳۹۸
https://doi.org/10.35050/JIPM010.2019.001

چکیده: واژه کوچک‌ترین واحد زبان است که دارای «صورت» و «معنا»ست. واژه ممکن است بیش‌از یک معنا داشته باشد و با توجه به کاربرد آن در بافت زبانی، معنای دقیق آن مشخص می‌شود. گردآوری تمام معانی یک واژه به‌صورت دستی کار بسیار پرزحمت و زمان‌بر است. افزون بر آن، ممکن است معانی واژه با گذشت زمان دچار تغییر شود؛ به این صورت که معانی موجود واژه کم‌کاربرد شده یا معانی جدید به آن اضافه شود. یکی‌از روش‌هایی که می‌توان برای تعیین معنای واژه استفاده کرد به‌کارگیری روش‌های رایانشی با توجه به بافت زبانی است.

در پژوهش حاضر تلاش می‌شود با ارائهٔ یک الگوریتم محاسباتی، معانی واژه‌های هم‌نگارهٔ فارسی با توجه به بافت زبانی به‌صورت خودکار و بدون نیاز به ناظر انسانی تعیین شود. برای رسیدن به این هدف، از روش تعبیهٔ معنای واژه در یک مدل فضای برداری استفاده می‌شود. برای ساخت بردار واژه، از یک رویکرد مبتنی بر شبکهٔ عصبی استفاده می‌شود تا اطلاعات بافت جمله به خوبی در بردار واژه گنجانده شود. در گام بعدی مدل پیشنهادی، برای ساخت بردار متن و تعیین معنای واژه، دو حالت جمله‌بنیان و بافت‌بنیان معرفی می‌شود. در حالت جمله‌بنیان، تمام واژه‌های جمله‌ای که واژهٔ هدف در آن وجود دارد، در ساخت بردار نقش دارد؛ ولی در حالت بافت‌بنیان فقط تعداد محدودی از واژه‌های اطرافِ واژهٔ هدف برای ساخت بردار در نظر گرفته می‌شود. دو معیار ارزیابی درونی و برونی برای ارزیابی کارایی الگوریتم خوشه‌بندی به‌کار گرفته می‌شود. معیار ارزیابی درونی که محاسبهٔ مقدار تراکم داده در هر خوشه است برای دو حالت جمله‌بنیان و بافت‌بنیان محاسبه می‌گردد. ارزیابی برونی به دادهٔ استاندارد طلایی نیاز دارد که برای این هدف، یک مجموعه داده شامل ۲۰ واژهٔ هدف فارسی و تعداد ۱۰۰ جملهٔ نشانه‌گذاری‌شده برای هر یک از این واژه‌ها تهیه شده است. بر اساس نتایج به‌دست‌آمده از ارزیابی

درونی، تراکم خوشه‌ای حالت جمله‌بنیان با تفاوتی معنادار بالاتر از حالت بافت‌بنیان است. با در نظر گرفتن دو شاخص V و F در ارزیابی برونی، مدل بافت‌بنیان به‌صورتی معنادار کارایی بالاتری را نسبت به جمله‌بنیان و مدل‌های پایه به‌دست آورده است.

کلیدواژه‌ها: تعبیه معنایی واژه، خوشه‌بندی، یادگیری ماشین بی‌مربی، فضای برداری، پردازش زبان طبیعی، بازنمایی معنایی واژه، زبان فارسی

۱. مقدمه

زبان طبیعی در مقایسه با زبان برنامه‌نویسی[1] یا زبان صوری[2] یک پدیدهٔ چندوجهی است که در یک کانال ارتباطی برای انتقال مفهوم بین افراد یک جامعه به‌کار می‌رود. «دوسوسور» زبان را متشکل از دو وجه «صورت»[3] و «معنا»[4] می‌داند که «صورت» به‌واسطهٔ نظام آوایی یا نظام نوشتاری تجلی عینی می‌یابد و «معنا» در این کانال ارتباطی جنبهٔ انتزاعی دارد. همچنین وی بر دو محور همنشینی[5] و جانشینی[6] قائل است که آرایش واژه‌ها در کنار یکدیگر در محور همنشینی به ساخت جملهٔ معنادار می‌انجامد و واژه‌هایی که با یکدیگر رابطهٔ معنایی دارند، در محور جانشینی می‌توانند جانشین یکدیگر شوند (de Saussure 1916). چنانچه در این کانال مفهوم به‌درستی منتقل نشود، پذیرنده نمی‌تواند آن را درک نماید. به‌عبارات دیگر، مفهوم یک جمله در یک ارتباط موفق بین تولیدکننده و پذیرنده از ترکیب مفاهیم واژه‌های به‌کاررفته در آن جمله شکل می‌گیرد. می‌دانیم هر واژه حاوی معناست و در یک فرهنگ لغت تلاش می‌شود با کمک واژه‌های دیگر و ارائهٔ تعریف، معنای آن واژه مشخص گردد. گاهی رابطه بین «صورت» و «معنا» یک‌به‌یک بوده و تعیین معنای این دسته از واژه‌ها در این کانال ارتباطی بسیار ساده است؛ مانند «چاقو». گاهی این رابطه یک‌به‌یک نیست و در این شرایط موضوع ابهام مطرح می‌شود و به اختلال در این کانال ارتباطی می‌انجامد؛ مانند واژه «برداشت» که معنای آن در دو عبارت «برداشت محصول» و «برداشت از سخن» کاملاً متفاوت از یکدیگر است و رفع ابهام معنایی این واژه بدون بافت زبانی ممکن نیست.

ابهام یکی‌از ویژگی‌های زبان طبیعی است که آن را از زبان‌های غیرطبیعی متمایز می‌سازد. ابهام در لایه‌های مختلف زبانی وجود دارد که دو مورد آن بیشتر شناخته‌شده

1. programming language
2. formal language
3. form
4. meaning
5. syntagmatic axis
6. paradigmatic axis

است: الف) ابهام ساختاری؛ و ب) ابهام واژگانی. در ابهام ساختاری، توالی واژه‌ها در یک عبارت یا جمله بیش از یک مفهوم را در کانال ارتباطی بین تولیدکننده و دریافت‌کننده منتقل می‌کند؛ مانند «دو پسر و دختر جوان» که ابهام این عبارت، علاوه بر تعداد نفرات، یعنی دو پسر و دو دختر یا دو پسر و یک دختر، در ویژگی «جوانی» نیز هست؛ به این صورت که این ویژگی می‌تواند فقط برای «دختر» یا برای هم «دختر» و هم «پسر» باشد. ابهام واژگانی ممکن است در «صورت»، چه در نمود آوایی و چه نوشتاری، یا «معنا» تجلی یابد. از جمله دلایل ایجاد ابهام در زبان، وجود رابطهٔ چندمعنایی[1] و همنامی[2] بین واژه‌هاست که تشخیص این دو مشکل است (Lyons 1981, 146). در رابطهٔ چندمعنایی یک واژه دارای چند معنای مرتبط به‌هم است و این قبیل واژه‌ها به‌عنوان یک مدخل واژگانی در فرهنگ لغت تعریف می‌شوند؛ مانند واژه «آب» در مثال‌های «آب‌های آزاد» و «لیوان آب». در همنامی، دو یا چند واژه از نظر آوایی و نوشتاری همانند یکدیگر است. این قبیل واژه‌ها به‌صورت مدخل‌های جداگانه در فرهنگ لغت فهرست می‌شوند، مانند «شیر» که به مفهوم «حیوان»، «مایع خوراکی» یا «ابزار قطع و وصل» است. در مورد رابطه بین دو واژه یادآور می‌شود چنانچه دو واژه فقط از نظر آوایی شبیه یکدیگر باشند، آن دو واژه هم‌آوا[3] نامیده شده و چنانچه دو واژه فقط از نظر نوشتاری شبیه یکدیگر باشند، آن دو واژه هم‌نویسه[4] نامیده می‌شوند.

ممکن است نظام نوشتاری یک زبان بر هم‌نویسگی واژه‌ها تأثیرگذار باشد و بر چالش ابهام واژگانی بیفزاید. «قیومی، ممتازی و بی‌جن‌خان» تعدادی از چالش‌های خط فارسی در پردازش‌های رایانشی را ذکر کرده‌اند (Ghayoomi, Momtazi and Bijankhan 2010). «بی‌جن‌خان و مرادزاده» چهار دلیل را برای علت‌های هم‌نویسگی در فارسی ذکر کرده و یک طبقه‌بندی از هم‌نویسه‌های فارسی ارائه کرده‌اند (۱۳۸۳). «بی‌جن‌خان» و همکاران نیز طبقه‌بندی دیگری از هم‌نویسه‌های فارسی را بر اساس مقولهٔ دستوری‌شان ارائه کرده‌اند (Bijankhan et al. 2011). یکی‌از دلایل هم‌نویسگی در فارسی عدم بازنمایی واکه‌های کوتاه فتحه، ضمه و کسره در خط فارسی ذکر شده است. همچنین، «بی‌جن‌خان و علایی ابوذر» فاصله بین صورت نوشتاری واژه‌های فارسی و صورت آوایی را که با اصطلاح «عمق خط فارسی» معرفی شده، زیاد می‌دانند (۱۳۹۲). افزایش تعداد واژه‌های هم‌نویسه در واژگان

1. polysemy 2. homonymy 3. homophone 4. homograph

زبان فارسی موجب چالش در پردازش خودکار داده‌های این زبان می‌گردد. گاهی حتی با داشتن مصوت‌های کوتاه، ابهام در معنای واژه همچنان باقی است، مانند «مِهر» که در بافت‌های «مهر مادر»، «ماه مهر» و «روزنامهٔ مهر» معانی این واژه متفاوت است. گاهی ممکن است ترکیبی از علل چندمعنایی، همنامی و هم‌نویسگی در یک واژه دیده شود، مانند «مهر» و «کرم».

هم‌معنایی، هم‌نامی و هم‌نویسگی به‌عنوان علل ابهام واژگانی، تعاریف دقیق و مشخص دارد و این موارد در مطالعات زبان‌شناسی مستقل از یکدیگر در نظر گرفته می‌شود. ولی در روش‌های پردازش زبان طبیعی که از پیکرهٔ زبانی استفاده می‌شود، معمولاً ورودی الگوریتم صورت‌واژه در پیکره است و فقط صورت نوشتاری، صرف نظر از علت ابهام، پردازش می‌شود. در این پژوهش، مجموعهٔ حاصل از صورت‌واژه‌های چندمعنا، هم‌نام و هم‌نویسه را تحت عنوان «واژه‌های هم‌نگاره» معرفی می‌نماییم و تلاش می‌کنیم با استفاده از روش استنتاج استقرایی معنایی[1] به‌طور کاملاً خودکار، معانی واژه‌های هم‌نگارهٔ فارسی را در یک پیکرهٔ زبانی معین نماییم. ناگفته نماند که این روش با «ابهام‌زدایی معنایی واژه»[2] متفاوت است. مسئلهٔ مورد توجه در ابهام‌زدایی معنایی واژه این است که یک مجموعهٔ مشخص از معانی واژه از قبل وجود دارد و رایانه باید معنای واژهٔ هدف را از این مجموعهٔ محدود مشخص نماید؛ در حالی که در استنتاج استقرایی معنایی واژه این مجموعهٔ معنایی اولیه وجود ندارد و معنای واژه در بافت زبانی باید به‌صورت الگوریتمی تعیین شود.

ساختار مقالهٔ حاضر به این صورت است: در بخش ۲، بازنمایی معنایی واژه توضیح داده می‌شود. در بخش ۳، پیشینهٔ مطالعاتی در حوزهٔ تعبیهٔ واژگانی و کاربرد آن بررسی می‌گردد. در بخش ۴، الگوریتم پیشنهادی برای پردازش زبان فارسی و تعیین معانی واژه‌های هم‌نگاره معرفی می‌گردد. در بخش ۵، داده‌ها و ابزارهای مورد نیاز برای این پژوهش توضیح داده می‌شود. در بخش ۶، نتایج به‌دست‌آمده از الگوریتم پیاده‌سازی‌شده گزارش می‌شود و بالاخره با نتیجه‌گیری در بخش ۷، مقاله به پایان می‌رسد.

1. word sense induction
2. word sense disambiguation

۲. بازنمایی معنایی واژه

۲-۱. رابطهٔ بافت زبانی و معنا

محور جانشینی معرفی‌شده توسط «دو سوسور» با نظر «ویتگنشتاین» و «هریس» همسوست. «ویتگنشتاین» بیان می‌دارد «معنای واژه در کاربرد آن نهفته است» (Wittgenstein 1953). «هریس» نیز بر این اعتقاد است که واژه‌هایی که در یک بافت زبانی یکسان به‌کار می‌روند، این تمایل را دارند که از نظر معنایی به یکدیگر شبیه[1] باشند (Harris 1954). بنا بر نظر وی، معنای هر واژه منعکس‌کنندهٔ بافتی است که آن واژه در آن بافت به‌کار رفته است. نظر «هریس» در چارچوب «معناشناسی توزیعی»[2] قرار می‌گیرد که منجر به معرفی‌شدن «فرضیهٔ توزیعی»[3] شده است. بر اساس این فرضیه، واژه‌هایی که تمایل به حضور در یک بافت یکسان دارند، از نظر معنایی مشابه یکدیگرند. بنا بر نظر «هریس»، معنای واژه تأثیرپذیر از «بافت جایگاهی»[4] است که آن واژه در آن بافت ظاهر می‌شود (Harris 1954). در همین راستا، «فرث» می‌افزاید که با توجه به واژه‌های اطراف یک واژه می‌توان معنای یک واژه را مشخص کرد (Firth 1957). نتیجهٔ این نظرات آن است که بافت زبانی واژه، نقش بسیار مهمی در تعیین معنای یک واژه دارد. بر همین اساس، «میلر و چارلز» «فرضیهٔ بافتی قوی»[5] را مطرح می‌کنند که در آن شباهت دو واژه از نظر معنایی به اندازه‌ای است که بافت زبانی آن‌ها شبیه هم باشد (Miller and Charles 1991).

از مجموعهٔ نظرات مطرح‌شده می‌توان چنین نتیجه گرفت که واژه‌های «ماشین»، «خودرو» و «اتومبیل» در مثال‌های (۱) تا (۳) زیر به‌دلیل داشتنِ بافتِ مشابه، از نظر معنایی شبیه یکدیگر بوده و در محور جانشینی می‌توانند به‌جای یکدیگر به‌کار روند.

(۱) او ماشین را در پارکینگ پارک نمود.

(۲) او خودرو را در پارکینگ پارک نمود.

(۳) او اتومبیل را در پارکینگ پارک نمود.

در حوزهٔ پردازش زبان طبیعی نیاز است پدیدهٔ زبانی از نظر آماری و احتمالاتی قابل محاسبه باشد. بنابراین، میزان تشابه معنایی با استفاده از یک روش محاسباتی باید اندازه‌گیری گردد تا مقدار تشابه معنایی به‌صورت رقمی بیان شود.

---

1. similar
2. distributional semantics
3. distributional hypothesis
4. local context
5. strong contextual hypothesis

۲-۲. روش‌های بازنمایی معنایی

یکی از کارهای پردازشی داده زبانی، استخراج الگوهای زبانی به همراه آمار کاربردی آن‌هاست. این الگوها که همان بافت‌های جایگاهی کاربرد واژه‌هاست، از حجم زیاد داده استخراج می‌شود. از کنار هم قرار گرفتن این الگوها و انجام پاره‌ای محاسبات در یک الگوریتم می‌توان به مفاهیم نهفته دست یافت و از آن‌ها در تحلیل‌هایی با کاربری خاص استفاده نمود. برای نمایش اطلاعات بافتی در چارچوب «معناشناسی توزیعی»، «سونگ، وانگ و گیلدا» دو روش کلی را معرفی کرده‌اند: الف) استفاده از روش‌های مبتنی بر روش «بِیز»[1] که برای نمایش اطلاعات بافتی، رویکردهای مربوط به مدل‌سازی موضوع[2] را (Blei et al. 2003a) که کاملاً بی‌مربی[3] است، به‌کار می‌گیرد. ب) روش‌های مبتنی‌بر ویژگی[4] که برای نمایش اطلاعات بافتی از بازنمایی اطلاعات بافتی به‌صورت بردار استفاده می‌کند (Song, Wang and Gildea 2016). در این صورت، شکل داده‌ها از حالت اصلی صورت‌واژه خارج شده و بر اساس ویژگی‌های واژهٔ هدف به بردار تبدیل می‌گردد تا امکان انجام محاسبات مربوط به میزان تشابه بردارها میسر شود. انعطاف‌پذیری مدل‌های فضای برداری توجه پژوهشگران را به خود جلب کرده است تا در چارچوب استنتاج استقرایی معنایی، معانی واژه‌ها به‌دست آید.

مدل فضای برداری که از حوزهٔ بازیابی اطلاعات نشأت گرفته، با معناشناسی توزیعی همسوست تا اطلاعات مربوط به واژه و بافت آن واژه بازنمایی گردد. به‌عبارت دیگر، کاربرد فضای برداری سبب فشردگی اطلاعات مربوط به واژه‌ها و بافت کاربردی واژه‌ها می‌شود تا توزیع معنایی واژه‌ها را بیان کند. این شیوه ارائهٔ اطلاعاتِ مربوط به واژه، «تعبیهٔ واژه»[5] نامیده می‌شود (Mikolov et al. 2013). محاسبهٔ «فاصلهٔ هندسی»[6] بین بردارها یکی‌از راه‌های یافتن شباهت بین واژه‌هاست. در مثال‌های (۱) تا (۳) بالا، فاصلهٔ هندسی بردارهای «ماشین»، «خودرو» و «اتومبیل» به یکدیگر بسیار نزدیک است؛ بنابراین، فرض بر این خواهد بود که این واژه‌ها از نظر معنایی به یکدیگر شبیه هستند. برای محاسبهٔ فاصلهٔ بردارها، معمولاً از معیارهای محاسباتی، مانند فاصلهٔ اقلیدوسی[7] و فاصلهٔ کسینوسی[8] استفاده می‌شود (Jurafsky and Martin 2018).

---

1. Bayes
2. topic modeling
3. unsupervised
4. feature
5. word embedding
6. geometric distance
7. Euclidean distance
8. Cosine distance

۲-۳. روش‌های مدل‌سازی بافت در بازنمایی معنایی

خلاصه‌سازی حجم زیادی از اطلاعات در یک بردار خطی سبب شده که استفاده از تعبیهٔ واژه مورد توجه قرار گیرد. روش‌های بازنمایی معنایی باید به‌گونه‌ای در محیط رایانه مدل‌سازی شود. از این رو، دو روش برای فشرده‌سازی اطلاعات بافت جایگاهی معرفی شده است: الف) روش مبتنی‌بر تجزیهٔ ماتریس که به «بازنمایی بردار جهانی»[1] معروف است (Pennington, Socher and Manning 2014)؛ ب) روش‌های مبتنی‌بر شبکهٔ عصبی که در آن از «مدل‌های زبانی عصبی»[2] استفاده می‌شود. یکی‌از ویژگی‌های «مدل‌های زبانی عصبی» این است که علاوه بر آموزش یک مدل زبانی بر مبنای شبکهٔ عصبی، ساختاری را برای نگاشت[3] واژه‌ها به فضای برداری فراهم می‌نماید. یادگیری این نگاشت به‌واسطهٔ بهینه‌سازیِ یک تابعِ هدف صورت می‌گیرد. «میکولف» و همکاران او این تابع هدف را به دو صورت تعریف می‌کنند: الف) یادگیری یک بردار برای واژهٔ هدف که بتواند بردار واژه‌های بافت را پیش‌بینی کند. این شیوه به مدل «پرش‌نگاشت پیوسته»[4] معروف است و برای بازنمایی واژه مورد استفاده قرار می‌گیرد. این مدل بازنمایی، به‌دنبال بهینه‌سازیِ توانایی بردارِ هر واژه در پیش‌بینیِ بردارِ واژه‌های اطراف آن است. ب) یادگیری بردارهای واژه‌های بافت که بتواند بردار واژهٔ هدف را پیش‌بینی کند که به مدل «کیسه‌واژه پیوسته»[5] معروف است و به‌دنبال بهینه‌سازیِ توانایی بردارهای اطراف هر واژهٔ هدف در پیش‌بینیِ بردارِ واژهٔ هدف است (Mikolov et al. 2013).

۳. پیشینهٔ مطالعاتی تشخیص معنای واژه با استفاده از روش‌های بازنمایی معنایی

پژوهش «پنتل و لین» جزء اولین پژوهش‌های انجام‌شده با هدف استفاده از خوشه‌بندی برای تشخیص معنای واژه است. بنا بر نظر آن‌ها، واژه‌هایی که در بافت‌های متنی مشابه، بافت‌های نحوی مشابه و اسناد مشابه به‌کار می‌روند، از نظر معنایی مشابه هم هستند (Pantel and Lin 2002). آن‌ها در این پژوهش به‌دنبال یافتنِ معانیِ جدید یا نادرِ واژه در متن، با کمک «تشابه توزیعی» هستند. برای این هدف، الگوریتم Lin (1998) به‌کار گرفته‌شده که در آن از روابط نحوی برای خوشه‌بندی استفاده شده است.

«وندوکرویز و آپیدیناکی» و «لاوو» و همکاران از مدل‌سازی موضوع برای تشخیص

1. Global Vector (GLOVE)
2. neural language model
3. mapping
4. continuous Skip-gram (Skip-gram)
5. Continuous Bag Of Words (CBOW)

چندمعنایی و یافتن معانی جدید استفاده کرده‌اند. آن‌ها با کمک مدل‌سازی موضوع، ویژگی‌های مربوط به موضوع را به بردار تبدیل کرده و سپس، بردار به‌دست‌آمده را خوشه‌بندی کرده‌اند (Van de Cruys and Apidianaki 2011؛ Lau et al. 2012).

«هوانگ» و همکاران یک مدل زبانی بر اساس شبکهٔ عصبی معرفی کرده‌اند که در این مدل، بافت جایگاهی و گستردهٔ واژه در یک سند به‌کار می‌رود. در پژوهش آن‌ها، برای بازنمایی بهتر معنای واژه، از مدل فضای برداری استفاده شده است (Huang et al. 2012). در این مدل برداری، به‌واسطهٔ اطلاعات بافت جایگاهی، علاوه بر معنا، اطلاعات نحوی واژه‌ها نیز به‌صورت ضمنی حفظ می‌شود. ویژگی پژوهش آن‌ها این است که بر اساس فاصلهٔ بردارها می‌توان میزان تشابه معنایی واژه‌ها را سنجید. در انجام این پژوهش، از پیکرهٔ «ویکی‌پدیای»[1] ۲۰۱۰ انگلیسی برای آموزش مدل و ساخت بردار استفاده شده است (Shaul and Westbury 2010). به هنگام استخراج بافت جایگاهی واژه برای ساخت بردار، ۱۰ واژهٔ همسایه، ۵ واژه در سمت چپ و ۵ واژه در سمت راست استخراج شده است. برای این منظور، در یک ماتریس دو بُعدی واژه، تعداد هم‌رخدادی[2] واژهٔ هدف با ۱۰ واژهٔ بافت جایگاهی محاسبه شده و از مدل «بسامد واژه ـ قلب بسامد سند»[3] (Salton et al. 1975) برای وزن‌دهی به آمار این هم‌رخدادی استفاده شده است. بردار تهیه‌شده برای هر واژه دارای ابعاد ۵۰ بُعدی است. الگوریتم خوشه‌بندی به‌کاررفته در این پژوهش، الگوریتم کِی-مینز[4] (MacQueen 1967) است. با توجه به این‌که این الگوریتم نیاز به پارامتر تعداد خوشه دارد، در این پژوهش این مقدار به‌صورت ثابت عدد ۱۰ در نظر گرفته شده است. بدیهی است تعیین این پارامتر به‌صورت ثابت و یکسان برای تمام واژه‌ها از جملهٔ نقاط ضعف این پژوهش محسوب می‌شود.

«نیلاکانتان» و همکاران از مدل فضای برداری برای استنتاج استقرایی معنایی واژه استفاده کرده‌اند. در این پژوهش، از روش‌های توزیعی برای تعبیهٔ واژه استفاده شده و مدل پرش‌نگاشت برای ساخت بردار به‌کار برده شده است (Neelakantan et al. 2014). مدل پرش‌نگاشت به ازای ظهور هر صورت‌واژه، بردار آن واژه را به‌روز می‌کند؛ در حالی که در مدل توسعه‌داده‌شده توسط آن‌ها، ابتدا معانی نزدیکِ واژهٔ هدف با توجه

1. Wikipedia
2. co-occurrence
3. Term Frequency-Inverse Document Frequency (TF-IDF)
4. K-means

به بافت جست‌وجو می‌شود و سپس، بردار معنایی واژه به‌روز می‌گردد. بنابراین، در مدل توسعه‌داده‌شده، بافت جایگاهی واژه بر ساختِ بردارِ واژه مؤثر است؛ و نتیجهٔ به‌دست‌آمده حاوی چندین بردار برای یک واژه در بافت‌های مختلف است که این شیوه «تعبیهٔ معنایی»[1] نامیده می‌شود.

«لی و جورافسکی» به‌جای استفاده از الگوریتم‌های خوشه‌بندی پارامتری، مانند کِی-مینز، از یک الگوریتم غیرپارامتری بنام پردازش رستوران چینی[2] (Blei et al. 2003b) استفاده کرده‌اند (Li and Jurafsky 2015). این الگوریتم به این صورت عمل می‌کند که آیا باید معنای جدید برای یک واژه در نظر گرفته شود یا این‌که این معنا مربوط به معانی قبلی است. برای ساخت بردار هر واژه، از «ویکی‌پدیای» ۲۰۱۴ انگلیسی استفاده شده است و تعداد معانی دویست‌هزار واژهٔ پربسامد انگلیسی از این پیکره به‌صورت استنتاج استقرایی به‌دست آمده است. برای ساخت این بردار، ۱۰ واژهٔ همسایه به‌عنوان بافت جایگاهی در نظر گرفته شده است. پس از معرفی الگوریتم بهبودیافته، این الگوریتم برای درک زبانی به‌صورت یک پیوستار مرحله‌ای، و انواع پردازش‌های زبانی، مانند تشخیص موجودیت‌های نامدار، برچسب‌دهی مقولهٔ دستوری، دسته‌بندی احساسات در سطح جمله، تحلیل احساسات، دسته‌بندی رابطهٔ معنایی و رابطهٔ معنایی جملات، مورد استفاده قرار گرفته است. با توجه به این‌که طبیعت این الگوریتم به این‌گونه است که «یک میزِ پُر، پُرتر می‌شود»، همیشه احتمال ایجاد یک معنای جدید برای یک واژه بسیار پایین است و معمولاً تعداد معانی به‌دست‌آمده توسط این الگوریتم کمتر از تعداد واقعی آن است.

«سونگ» و همکاران مدل فضای برداری را برای تعبیهٔ واژه و تعبیهٔ معنایی به‌کار برده‌اند (Song et al. 2016). Mikolov et al. (2013) برای ساخت بردارها، از ابزار Word2Vec استفاده کرده‌اند. سپس، بردار به‌دست‌آمده از واژه‌ها در مدل معرفی‌شده توسط (Neelakantan et al. 2014) با تعداد ثابت سه خوشه و مدل معرفی‌شده توسط Li and Jurafsky (2015) مورد استفاده قرار گرفته است. آن‌ها در انجام پژوهش خود از پیکرهٔ ویکی‌پدیای ۲۰۱۰ انگلیسی (Shaul and Westbury 2010) به‌عنوان داده آموزش[3] مدل و ساخت بردار استفاده شده است. در این پژوهشِ آن‌ها، داده همایش ۲۰۱۰ «ارزیابی

1. sense embedding
2. Chinese Restaurant Process
3. training data

معنایی»[1] که به SemEval2010 معروف است (Manandhar et al. 2010) به‌عنوان داده آزمون[2] مورد استفاده قرار گرفته است. در انتها، «سونگ» و همکاران مدل خود را با مدل‌های پایهٔ[3] معرفی‌شده در SemEval2010 مقایسه کرده‌اند. آن‌ها همچنین در این پژوهش، از الگوریتم کِی-مینز برای خوشه‌بندی استفاده کرده‌اند.

اکثر پژوهش‌های انجام‌شده برای استنتاج استقرایی معنای واژه، برای زبان انگلیسی انجام شده است و اکثر پژوهش‌های انجام‌گرفته در حوزهٔ تحلیل معنای واژهٔ فارسی، به موضوع «ابهام‌زدایی معنایی واژه» پرداخته‌اند که از یادگیری ماشینی با مربی[4] برای تشخیص معنای واژهٔ هدف از یک مجموعهٔ مشخص استفاده می‌شود؛ مانند، پژوهش‌های «سلطانی و فیلی» (۱۳۸۷)، «خسروی‌زاده و فارسی‌نژاد» (۱۳۹۱)، «شول و نورمندی‌پور» (۱۳۹۳)، «مسعودی و راحتی» (۱۳۹۴)، «ذوالفقاری‌کندری و موسوی‌میانگاه» (۱۳۹۴)، «کیانی‌نژاد، شیرازی و سدیدپور» (۱۳۹۵)، MosaviMiangah and DelavariKhalafi (2005)، Hamidi and Borji and ShiryGhidary (2007)، Sarrafzadeh et al. (2011a, and b)، Riahi and Sedghi (2012)، Rasekh, Sadreddini and Fakhrahmad (2014)، Rezapour et al. (2014)، Fakhrahmad, Sadreddini and Fakhrahmad et al. (2011, 2012)، ZolghadriJahromi (2014)، Mahmoodvand, and Hourali (2015)، Rekabsaz et al. (2016). در حالی که پژوهش حاضر با هدف استفاده از استنتاج استقرایی معنایی و کاربرد روش یادگیری ماشین بی‌مربی برای یافتن معانی واژه‌های هدف از پیکرهٔ زبانی انجام می‌پذیرد.

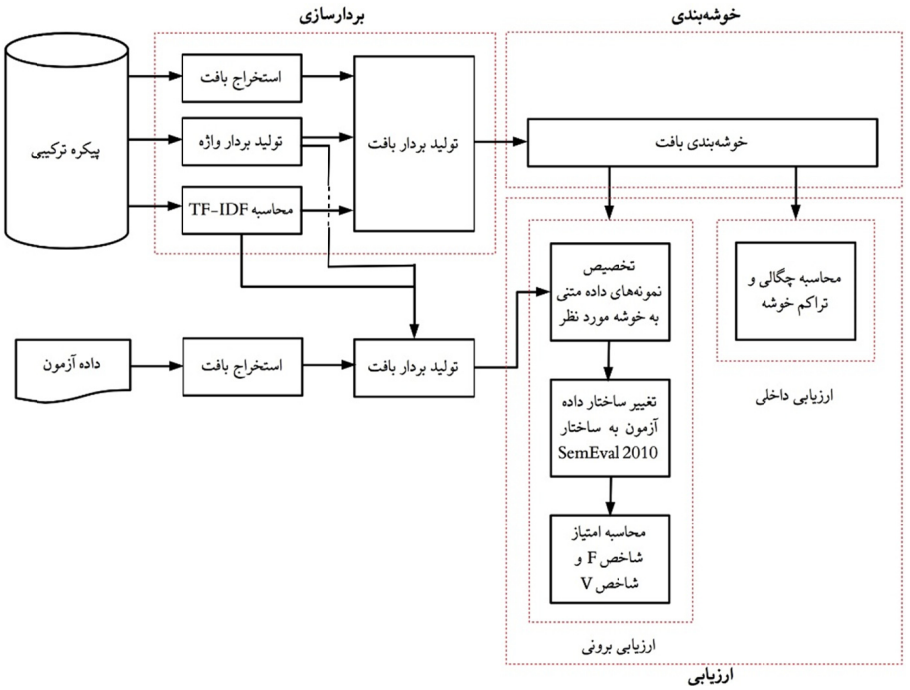
٤. معرفی مدل تشخیص معنای واژه

شکل ۱، مدل پیشنهادی تعیین معنای واژه‌های فارسی را نمایش می‌دهد. به‌طور خلاصه، روند اجرای این مدل پیشنهادی به این صورت است که ابتدا الگوریتم، یک پیکرهٔ زبانی بزرگ را به‌عنوان داده آموزش می‌پذیرد و بردار واژه‌ها را بر اساس این پیکره می‌سازد. سپس، بردارهای تهیه‌شده برای ساخت بردار جملات حاوی واژه‌های هم‌نگارهٔ هدف استفاده شده و به الگوریتم خوشه‌بندی داده می‌شود. در نهایت، خروجی الگوریتم خوشه‌بندی مورد ارزیابی قرار می‌گیرد. در ادامه، به شرح مفصل الگوریتم می‌پردازیم.

همان‌گونه که در شکل ۱، مشخص است، این مدل شامل ۳ قسمت اصلی است:

1. Semantic Evaluation (SemEval)
2. test data
3. baseline
4. supervised machine learning

(۱) تهیهٔ بردار از دادهٔ خامِ متنی، (۲) خوشه‌بندی[1] بردارها و (۳) ارزیابی نتایج. در این ساختار، دو مجموعه دادهٔ ورودی مورد نیاز است. مجموعه دادهٔ اول یک پیکرهٔ خامِ زبانی با حجم زیاد است که برای ساخت دو نوع بردار مورد استفاده قرار می‌گیرد. مجموعهٔ دادهٔ دوم، یک پیکرهٔ کوچکِ نشانه‌گذاری‌شده است که به‌عنوان دادهٔ استاندارد طلایی[2] برای ارزیابی مدلِ معرفی‌شده مورد استفاده قرار می‌گیرد.



شکل ۱. مدل پیشنهادی تشخیص معانی واژه‌های فارسی

٤-۱. ساخت بردار

برای ساخت بردار از پیکرهٔ زبانی به دو مرحلهٔ ایجاد بردار نیاز داریم: (۱) ساخت بردار برای هر یک از واژه‌های پیکره؛ و (۲) ساخت بردار برای متن حاوی واژهٔ هدف که در ادامه بیشتر توضیح داده می‌شود.

1. clustering
2. gold standard

٤-١-١. ساخت بردار برای هر واژه در واژگان

روشی که برای ساخت بردار واژه‌ها در استنتاج استقرایی معنا معرفی می‌شود از شبکهٔ عصبی استفاده می‌کند. در این روش، هدف اصلی، ساخت یک مدل زبانی از پیکرهٔ ورودی است. علاوه بر این مدل زبانی، بردار واژه‌ها نیز ساخته می‌شود. علت استفاده از شبکهٔ عصبی، کارایی بالای آن در مقایسه با سایر روش‌های محاسباتیِ مدل‌سازی برای ساختِ فضای برداری است. هر چقدر این فضای برداری حاوی اطلاعات دقیق باشد، مدل دقیق‌تری از زبان ساخته می‌شود.

سه ویژگی که در فشرده‌سازی اطلاعات بافت مربوط به واژه و ساخت بردار واژه در نظر گرفته می‌شود، از قرار زیر است: الف) ابعاد بردار؛ ب) تعداد واژه‌های بافت که در اطراف واژهٔ هدف ظاهر شده است؛ ج) نوع اطلاعات واژه‌های بافت در اطراف واژهٔ هدف. هرچه ابعاد بردار واژه بیشتر باشد، اطلاعات بیشتری در بردار جای می‌گیرد و البته، سبب افزایش پیچیدگی محاسبات می‌گردد. بررسی پژوهش‌های انجام‌شدهٔ مرتبط نشان می‌دهد که تعداد ۳۰۰ بُعد به‌عنوان تعداد ابعاد بردار بهینه است. برای مشخص‌شدن بافت نیز می‌توان تعداد ۸ واژهٔ اطراف واژهٔ هدف (۴ واژه در سمت راست و ۴ واژه در سمت چپ واژهٔ هدف) را تعریف نمود. در پژوهش حاضر، فقط از صورت‌واژه به‌عنوان اطلاعات واژه‌های بافت در ساخت بردار استفاده می‌شود و استفاده از سایر اطلاعات زبان‌شناختی، مانند رابطهٔ نحوی یا بنِ واژه، و بررسی تأثیر ساخت این نوع بردارها، به پژوهش‌های آتی محول می‌شود.

٤-١-٢. ساخت بردار برای متن حاوی واژه‌های هدف

پس‌از ساخت بردار واژه‌های موجود در واژگان، باید برای هر یک از متونی که واژهٔ هدف در آن ظاهر شده نیز بردار ایجاد گردد تا بتوان با خوشه‌بندی این بردارها، به استنتاج معنای واژهٔ هدف در متن پرداخت.

در مرحلهٔ ساختِ بردار از متن، به تعریف مناسب بافت نیاز داریم. در این بخش از معماری، منظور از بافت این است که برای قضاوت در مورد معنای یک واژه، به چه میزان اطلاعات در مورد واژه‌های اطراف واژهٔ هدف نیاز داریم. برای این منظور می‌توان تمام جمله‌ای را که واژهٔ هدف در آن ظاهر شده، به‌عنوان بافت متنی در نظر گرفت. این مدل را «حالت جمله‌بنیان» می‌نامیم و یا باید این بافت را به تعدادی از واژه‌های اطراف واژهٔ هدف محدود نمود که این مدل را «حالت بافت‌بنیان» می‌نامیم. در پژوهش حاضر، هر دو

حالت فوق بررسی و مقایسه خواهد شد. لازم به ذکر است که در حالت بافت‌بنیان، تعداد محدود ۸ واژه، همانند تعداد واژه‌های دخیل در ساخت بردار، به‌عنوان بافت جایگاهی واژهٔ هدف در نظر گرفته می‌شود. بعد از مشخص نمودن بافت متنی که برای نمایش واژهٔ هدف مورد استفاده قرار می‌گیرد، می‌توان با میانگین‌گیری از بردار تک‌تک واژه‌های موجود در آن بافت متنی، به بردار جمله و متن دست یافت.

می‌دانیم تمام واژه‌های زبان حاوی محتوا نیستند و مجموعه‌ای از واژه‌های زبان فقط نقش دستوری برعهده دارند؛ مانند حروف اضافه، حروف ربط و مانند آن. این دسته از واژه‌ها که بار محتوایی بسیار کمی دارند به ایست‌واژه[1] معروف هستند. در ساخت عادی بردار متن، تمام واژه‌ها اعم از واژهٔ محتوایی یا نقشی دخیل هستند؛ ولی می‌توان با روش وزن‌دهی به واژه‌ها، بر میزان اهمیت واژه‌های محتوایی افزود و از اهمیت واژه‌های دستوری کاست. برای این هدف از روش وزن‌دهی TF-IDF استفاده می‌شود. برای این منظور، ابتدا بسامد تمام واژه‌ها از متن هدف و پیکرهٔ زبانی استخراج شده و مقدار TF-IDF برای تمامی واژه‌ها محاسبه می‌گردد. با داشتن بردار واژه‌ها و همچنین مقدار وزن حاصل از TF-IDF و تعریف بافت مورد استفاده برای ساختِ بردارِ متن می‌توان میانگین وزن‌دار واژه‌های بافت اطراف واژهٔ هدف را محاسبه نمود و به‌عنوان بردار بافت واژهٔ هدف مورد استفاده قرار داد که در این پژوهش از همین شیوه استفاده شده است.

٤-۲. الگوریتم خوشه‌بندی

تعیین معانی واژه با کمک روش‌های استنتاج استقرایی معنای واژه بر اساس نظر Harris (1954) معرفی شده است. در استنتاج استقرایی معنای واژه می‌توان از الگوریتم‌های خوشه‌بندی که بی‌مربی بوده و مناسب این کار است، استفاده نمود. دادهٔ ورودی الگوریتم خوشه‌بندی به‌صورت برداری حاصل از بافتِ واژه است که روش ساخت آن در بخش ۴-۱-۲ توضیح داده شد. هر چقدر بردار ورودی دقیق‌تر باشد، خوشه‌بندی داده با دقت بیشتری انجام می‌پذیرد.

در پژوهش حاضر، تعداد معانی واژه از قبل نامشخص است. در مدل پیشنهادی برای تعیین معنای واژه با استفاده از روش استنتاج استقرایی، الگوریتمِ خوشه‌بندیِ سلسله‌مراتبی[2]

1. stop-word
2. hierarchical method

برای دو حالت جمله‌بنیان و بافت‌بنیان به‌کار می‌رود. در این پژوهش، الگوریتم جداگرایانه[1] که جزء خوشه‌بندی سلسله‌مراتبی است، به‌کار برده می‌شود. در شروع الگوریتم جداگرایانه، ابتدا تمام داده‌ها در یک خوشه قرار می‌گیرد. سپس، این الگوریتم به‌صورت یک حلقهٔ تکراری تلاش می‌کند در هر مرحله، یکی از خوشه‌های موجود در داده را به دو خوشه تفکیک کند. به‌دلیل تکرارپذیری این الگوریتم خوشه‌بندی، این چرخه باید در مرحله‌ای متوقف گردد. معیاری که ما در این مدل برای توقف این الگوریتم معرفی کرده‌ایم، محاسبهٔ مقدار تراکم[2] و تیرگی داده در هر مرحله از خوشه‌بندی است. تا زمانی‌که روند تغییر تراکم و تیرگی در مراحل متوالی افزایشی باشد، فرایند خوشه‌بندی ادامه پیدا می‌کند و با کاهش روند تیرگی، فرایند خوشه‌بندی متوقف می‌شود. تعداد خوشه‌ها در آخرین مرحلهٔ خوشه‌بندی، به‌عنوان تعداد معنای واژهٔ هدف ثبت می‌گردد.

٤-٣. ارزیابی

یکی دیگر از قسمت‌های مدل معرفی‌شده، ارزیابی است. ارزیابی کمّیِ مدل معرفی‌شده به دو دسته ارزیابی درونی و برونی تقسیم می‌گردد. منظور از ارزیابی درونی این است که بتوان بدون نیاز به دادهٔ استاندارد طلایی، نتیجهٔ حاصل از استنتاج استقرایی معنای واژهٔ هدف را ارزیابی کرد. یکی از معیارهای مطرح برای ارزیابی درونی، معیار تراکم و تیرگی خوشه‌بندی است. استفاده از این معیار نشان می‌دهد که خروجی حاصل از خوشه‌بندی به چه میزان انسجام رسیده است که می‌تواند معیار مناسبی برای بررسی عملکرد مدل باشد. بنابراین، برای ارزیابی درونی خوشه‌بندی، مقدار تراکم هر خوشه محاسبه می‌گردد. منظور از ارزیابی برونی این است که بتوان با استفاده از دادهٔ طلایی استاندارد، نتیجهٔ حاصل از مدل معرفی‌شده را بررسی کرد. برای این هدف باید تعدادی متن به‌عنوان دادهٔ استاندارد طلایی نشانه‌گذاری شود. تهیهٔ این دادهٔ استاندارد باید توسط فردی خبره انجام پذیرد. از آنجا که این داده برای زبان فارسی موجود نیست، در راستای پژوهش حاضر، این مجموعهٔ داده برای زبان فارسی تهیه شده است.

با در اختیار داشتن دادهٔ آزمون، باید فرایندی مشابه آن‌چه که برای ساختِ بردارِ بافتِ واژهٔ هدف در بخش ۴-۱-۲ توضیح داده شد، برای این داده نیز طی شود تا بافت واژه‌های

1. divisive
2. density

هدف در دادهٔ آزمون نیز به بردار تبدیل گردد و برای ارزیابی برونی مورد استفاده قرار گیرد. پس از خوشه‌بندی دادهٔ آزمون توسط مدل، امکان ارزیابی خروجی مدل بر اساس اطلاعات صحیح در دادهٔ استاندارد طلایی میسر می‌گردد. در این پژوهش، برای ارزیابی برونی، شاخص F (Van Rijsbergen 1979) و شاخص V (Rosenberg and Hirschberg 2007) که معیارهای استاندارد ارزیابی خوشه‌بندی هستند، مورد استفاده قرار می‌گیرد. تفاوت این دو شاخص در آن است که شاخص V همگنی[1] و تمامیت[2] را دربرمی‌گیرد، در حالی که شاخص F فقط همگنی را پوشش می‌دهد.[3]

علوم تجربی بر شواهد تجربی و انجام آزمایش استوار است. برای این‌که بتوان برتری یک ایده را سنجید باید این ایده را نسبت به یک بستر کنترل‌شده به‌صورت تجربی ارزیابی نمود. این بستر کنترل‌شده مدل پایه[4] نامیده می‌شود. در این پژوهش، دو مدل پایهٔ ابتدایی مورد استفاده قرار می‌گیرد: الف) «یک معنا به‌ازای هر خوشه»[5]؛ و ب) «پربسامدترین معنا»[6]. در معیار پایهٔ اول، هر واژه در یک خوشهٔ مجزا قرار می‌گیرد. بنابراین، به تعداد داده‌های آزمون برای واژهٔ هدف، خوشه وجود دارد. در معیار پایهٔ دوم، تمام داده‌ها در خوشه‌ای که بالاترین بسامد را دارد، قرار می‌گیرد. روش کار در این مدل پایه به این صورت است که ابتدا الگوریتم کار خوشه‌بندی را انجام می‌دهد و هر جمله در خوشه‌ای قرار می‌گیرد و برای تشخیص خوشه‌ها، به هر خوشه یک اندیس[7] داده می‌شود. تعداد مواردی که در هر خوشه قرار می‌گیرد، یکسان نیست. از بین خوشه‌ها، خوشه‌ای که بالاترین بسامد را دارد، انتخاب می‌شود و اندیس واژه‌های هدف در دیگر خوشه‌ها با اندیس خوشهٔ پربسامد در تمام دادهٔ آزمون جایگزین می‌شود.

علاوه بر این دو، مدل پیشنهادی با «آخرین وضعیت روز»[8] نیز مقایسه می‌شود. بر اساس پژوهش Song et al. (2016)، الگوریتم خوشه‌بندی کی-مینز با تعداد **۳** خوشه به‌عنوان مدل پایهٔ «آخرین وضعیت روز» برای فارسی معرفی می‌شود.

---

1. homogeneity
2. completeness
۳. به‌دلیل کمبود فضا از ذکر فرمول‌های محاسبه این دو شاخص اجتناب شده است.
4. baseline
5. one sense per cluster (1S1C)
6. the most frequent sense
7. index
8. state-of-the-art

۵. داده‌ها و ابزارهای مورد نیاز

۵-۱. داده‌های پژوهش

از آنجا که در این پژوهش یک مدل محاسباتی برای زبان فارسی پیشنهاد می‌گردد، به دو دستهٔ دادهٔ زبان فارسی نیاز است: دستهٔ اول دادهٔ آموزش برای ساخت بردارهای واژگان و دستهٔ دوم دادهٔ آزمون است. پیکرهٔ زبانی بزرگی که در پژوهش حاضر به‌عنوان دادهٔ آموزش برای ساختِ بردارِ واژه‌های فارسی به‌کار می‌رود، از ترکیب چند مجموعه داده شکل گرفته است که خلاصهٔ آن در جدول ۱، ارائه شده است. پایگاه داده‌های زبان فارسی Assi (1997)، پیکرهٔ بی‌جن‌خان (۱۳۸۳)، پیکرهٔ همشهری AleAhmad et al. (2009) حاصل از آرشیو روزنامهٔ همشهریِ سال ۱۳۷۵ تا ۱۳۸۵، پیکرهٔ روزنامه‌ای حاصل از آرشیو برخط چندین روزنامه و ویکی‌پدیای فارسی[1] متشکل از ۳۶۱۴۷۹ مقاله از آرشیو ۲۰۱۶ وبگاه شکل گرفته است. مجموع این پیکرهٔ ترکیبی که برای ساخت بردار واژه‌ها به‌کار می‌رود، در حدود ۵۳۹ میلیون واژه است.

جدول ۱. مجموعه داده‌های به‌کاررفته در پژوهش حاضر

نام پیکره	نام منبع گردآوری	تعداد واژه‌های با تکرار	تعداد واژه‌های بدون تکرار
پایگاه داده‌های زبان فارسی	پژوهشگاه علوم انسانی و مطالعات فرهنگی	۲۳,۸۴۸,۶۵۵	۳۰۷,۷۲۶
پیکرهٔ همشهری	گروه پژوهشی پایگاه داده، دانشگاه تهران	۱۵۷,۸۴۱,۱۲۳	۶۰۸,۵۰۳
پیکرهٔ روزنامه‌ای	پژوهشگاه علوم انسانی و مطالعات فرهنگی	۳۰۱,۱۸۶,۲۷۷	۱,۳۲۴,۳۱۷
پیکرهٔ بی‌جن‌خان	گروه پژوهشی پایگاه داده، دانشگاه تهران	۲,۶۰۲,۵۳۶	۷۷,۱۴۳
ویکی‌پدیای فارسی	وبگاه ویکی‌پدیا	۸۰,۹۹۵,۷۴۳	۹۵۶,۶۵۵
تعداد کل		۵۳۸,۵۸۶,۴۸۷	۱,۹۵۷,۵۴۱

علاوه بر پیکرهٔ «همشهری»، حجم اعظم این پیکرهٔ ترکیبی متشکل از «پیکرهٔ روزنامه‌ای» است که از طریق خزش[2] آرشیو چندین روزنامه در بازهٔ زمانی ۱۳۸۰ تا ۱۳۹۶ گردآوری شده و به‌صورت یک پیکره ساماندهی شده است. برای استفاده از این پیکره باید یکدستی در داده رعایت شود و بر اساس یک استاندارد مشخص، این حجمِ داده

1. https://archive.org/details/fawiki-20160720 (آخرین دسترسی در مردادماه ۱۳۹۸)
2. crawl

ساماندهی شود. برای این هدف از مدل معرفی‌شده توسط «قیومی» (۱۳۹۸) در «نرمال‌سازی پیکره»[1] و «واحدسازی پیکره»[2] استفاده می‌شود.

دستهٔ دوم داده‌ای که به آن نیاز است، یک پیکره با حجم کوچک به‌عنوان دادهٔ آزمون است که برای ارزیابی مدلِ معرفی‌شده استفاده می‌شود. در پژوهش حاضر، ۲۰ واژهٔ هم‌نگاره به‌عنوان واژهٔ هدف، صرف نظر از علت ابهام و مقولهٔ دستوری آن واژه‌ها از «فارس‌نِت» (Shamsfard et al. 2010) انتخاب شده و جملات شاهد مربوط به این واژه‌ها از پایگاه داده‌های زبان فارسی (Assi 1997) استخراج شده است. در جدول ۲، تعداد معانی استخراج‌شده از «فارس‌نِت» برای واژه‌های هدف گزارش شده است. با بررسی معانی تعریف‌شده برای واژه‌های هدف در «فارس‌نِت» دریافتیم که بعضی از معانی واژه‌ها تکراری است و یا می‌توان بعضی از معانی واژه‌ها را به‌صورت کلی‌تر عنوان نمود. همچنین، در جملات شاهدِ گردآوری‌شده، معانی‌ای را یافتیم که در این مجموعهٔ معانی تعریف نشده بود و این موارد به این مجموعه اضافه شد. تغییرات تعداد معانی واژه‌های هدف در دادهٔ آزمون در جدول ۲، قابل مشاهده است.

جدول ۲. تعداد معانی استخراج‌شده از «فارس‌نِت» برای واژه‌های هدف

واژهٔ هدف	تعداد معانی		واژهٔ هدف	تعداد معانی	
	فارس‌نت	پژوهش حاضر		فارس‌نت	پژوهش حاضر
برداشت	۴	۵	سیر	۷	۶
پروانه	۴	۴	شیر	۱۳	۷
تار	۹	۸	کرم	۱۲	۹
تن	۸	۶	کره	۷	۵
تند	۷	۵	گرد	۶	۶
تیز	۵	۵	گل	۱۱	۶
روان	۵	۵	ملک	۶	۶
ریش	۲	۶	مهر	۱۰	۸
سبک	۸	۳	نبرد	۳	۸
سر	۲۴	۱۲	نشسته	۳	۳

1. normalization 2. tokenization

با در نظر داشتن مجموعهٔ جدید معانی در این پژوهش، ۱۰۰ جمله برای هر واژهٔ هدف توسط نیروی انسانی خبره به‌صورت دستی تحلیل شده و برای ساماندهی اطلاعات نشانه‌گذاری معنایی واژه‌های هدف فارسی، از استاندارد معرفی‌شده در SemEval2010 استفاده شده است. لازم به ذکر است که دادهٔ SemEval2010 حاوی ۸۰۹۱۵ نمونه جمله برای ۱۰۰ واژهٔ هدف انگلیسی، ۵۰ واژه با مقوله دستوری اسم و ۵۰ واژه با مقوله دستوری فعل است (Manandhar et al. 2010).

۵-۳. ابزارهای پژوهش

در این پژوهش از زبان برنامه‌نویسی «پایتون»[1] و مجموعه‌ای از کتابخانه‌ها مانند Gensim برای ساخت بردار واژه‌ها و کتابخانه‌های Cluster و Sklearn برای الگوریتم‌های خوشه‌بندی مورد نیاز در مدل معرفی‌شده استفاده شده است. همچنین، برای ارزیابی مدل معرفی‌شده در پژوهش حاضر، از ابزارهای ارائه‌شده در SemEval2010 برای ارزیابی شاخص V و شاخص F استفاده می‌شود.[2]

٦. نتایج حاصل از آزمایش‌های مدل پیشنهادشده

٦-١. ارزیابی برونی

در جدول ۳، نتایج به‌دست‌آمده از دو دسته از آزمایش‌های مدل‌های پایه در ارزیابی برونی بر اساس دو معیار شاخص V و شاخص F گزارش شده است. همان‌گونه که از نتایج به‌دست‌آمده از مدل‌های پایهٔ ابتدایی مشخص است، کارایی مدل «یک معنا به ازای هر خوشه» بر اساس شاخص V بسیار عالی و بر اساس شاخص F بسیار ضعیف است. همچنین، کارایی مدل «پربسامدترین معنا» بر اساس شاخص F بسیار عالی و بر اساس شاخص V بسیار ضعیف است. با توجه به این‌که هر دو این شاخص‌ها در ارزیابی از اهمیت به‌سزایی برخوردارند، هیچ‌یک از این دو مدل ایده‌آل نبوده و مدلی که بتواند هر دو شاخص را در حد قابل قبول بالا نگه دارد، از ارزش بیشتری برخوردار است. در نتیجه، ازمیان مدل‌های پایه در جدول ۳، اگرچه مدل «آخرین وضعیت روز» از نظر امتیاز شاخص V پایین‌تر از مدل «یک معنا به ازای هر خوشه» و از نظر امتیاز شاخص F پایین‌تر از مدل «پربسامدترین

1. python
2. https://www.cs.york.ac.uk/semeval2010_WSI/datasets.html

معنا»ست، با لحاظ‌کردن هر دو معیار، بهتر از دو مدل پایهٔ دیگر عمل کرده است.

جدول ۳. نتایج به‌دست‌آمده از مدل‌های پایه برای داده فارسی

مدل	حالت	شاخص V (درصد)	شاخص F (درصد)
پایهٔ ابتدایی	یک معنا به‌ازای هر خوشه	۳۷/۳۰	۰/۰۷
	پربسامدترین معنا	۰/۱۰	۵۹/۵۱
آخرین وضعیت روز	گسستی-۳-جمله‌بنیان	۲۶/۷۰	۵۱/۸۴

در جدول ۴، نتایج به‌دست‌آمده از دو حالت مدل پیشنهادی گزارش شده است. از میان حالت‌های مدل پیشنهادی، حالت بافت‌بنیان با استفاده از الگوریتم سلسله‌مراتبی، در هر دو شاخص V و F نسبت به حالت جمله‌بنیان امتیاز بالاتر به‌دست آورده است. لازم به ذکر است که بر اساس «آزمون تی دو جهته»[1]، تفاوت ۳/۶ درصدی شاخص V در حالت بافت‌بنیان نسبت به حالت جمله‌بنیان معنادار است ($p < ۰/۰۵$). این نتیجه بیان می‌کند که بافت جایگاهی واژهٔ فارسی در تعیین معنای آن واژه بسیار حائز اهمیت است.

جدول ٤. نتایج به‌دست‌آمده از مدل‌های پیشنهادی برای دادهٔ فارسی

مدل	حالت	شاخص V (درصد)	شاخص F (درصد)
مدل پیشنهادی سلسله‌مراتبی بر اساس تیرگی	جمله‌بنیان	۲۴/۰۰	۵۶/۷۹
	بافت‌بنیان	۲۹/۶۰	۵۸/۹۴

بر اساس نتایج به‌دست‌آمده از دو جدول ۳ و ۴، حالت بافت‌بنیان نسبت به حالت جمله‌بنیان و مدل‌های پایه «پربسامدترین معنا» و «آخرین وضعیت روز»، بالاترین امتیاز شاخص V را به‌دست آورده است؛ ولی این مدل نتوانسته مدل پایهٔ ابتدایی «یک معنا به ازای هر خوشه» را شکست دهد. همان‌گونه که ذکر شد، اگرچه شاخص V از اهمیت بیشتری نسبت به شاخص F برخوردار است، برای مقایسه نباید کاملاً ارزش شاخص F را نادیده گرفت. بر این اساس، از مقایسهٔ مدل پیشنهادی در پژوهش حاضر و مدل پایهٔ «آخرین وضعیت روز» این نتیجه به دست می‌آید که حالت-بافت‌بنیان بر اساس شاخص‌های V

1. two tailed t-test

و F علاوه بر حالت جمله‌بنیان، نسبت به مدل پایهٔ «آخرین وضعیت روز» کارایی بهتری دارد؛ در حالی که حالت جمله‌بنیان فقط در شاخص F کارایی بهتری نسبت به این مدل پایه به‌دست آورده است.

۶-۲. ارزیابی درونی

در ارزیابی درونی، مقدار تراکم و تیرگی داده در خوشه‌ها با استفاده از معیار تراکم در خوشه محاسبه می‌شود. نتایج به‌دست‌آمده از مقدار تراکم و تیرگی داده در خوشه‌های به‌دست‌آمده برای هر واژهٔ هدف، در جدول ۵ برای دو حالت جمله‌بنیان و بافت‌بنیان مدل پیشنهادی گزارش شده است. بر اساس نتایج حاصل از میانگین مقدار تراکم و تیرگی داده در خوشه‌های انتخاب‌شده واژه‌های هدف، حالت جمله‌بنیان مدل پیشنهادی، بالاترین مقدار تراکم را به‌دست آورده است. بر اساس آزمون تی دوسویه، تفاوت میزان تراکم و تیرگی در حالت سلسله‌مراتبی جمله‌بنیان نسبت به حالت بافت‌بنیان معنادار است ($p < ۰/۰۵$).

جدول ۵. میانگین مقدار تراکم داده در تعداد خوشه‌های انتخاب‌شده واژه‌های هدف فارسی

سلسله‌مراتبی-جمله‌بنیان	سلسله‌مراتبی-بافت‌بنیان
۰/۰۸۴۴	۰/۰۷۵۹

۷. نتیجه‌گیری

در پژوهش حاضر، به پردازش معنایی واژه و مشخص‌شدن تعداد معانی آن پرداخته شد. برای این منظور، یک مدل پردازشی مبتنی‌بر خوشه‌بندی و تعبیهٔ معنایی واژه‌ها با استفاده از شبکهٔ عصبی ارائه شد که بتواند کاملاً بدون نظارت انسان و با به‌کارگیری روش خوشه‌بندی داده، تعداد معانی واژه‌های هدف را مشخص کند. با توجه به این‌که برای ارزیابی مدل به دادهٔ استاندارد طلایی نیاز است، چنین داده‌ای برای فارسی موجود نبود. با در نظر داشتن ۲۰ واژهٔ هدف، در مجموع تعداد ۲۰۰۰ جمله، ۱۰۰ جمله برای هر واژه، به‌صورت دستی نشانه‌گذاری شد و بر اساس ساختار داده SemEval2010 ساماندهی شد. این داده‌ها برای ارزیابی برونی خوشه‌بندی به‌کار رفت. بر اساس نتایج به‌دست‌آمده، اساساً حالت‌های معرفی‌شدهٔ جمله‌بنیان و بافت‌بنیان با استفاده از الگوریتم سلسله‌مراتبی کارایی بهتری را بر اساس شاخص V در مقایسه با مدل‌های پایهٔ ابتدایی «پربسامدترین

معنا» و «آخرین وضعیت روز» به‌دست آورد. این نتیجه بیانگر آن است که استفاده از روش ساخت مدل فضای برداری مبتنی بر شبکهٔ عصبی، بسیار خوب توانسته اطلاعات مربوط به واژه را از بافت به‌دست آورد. علاوه بر مقایسهٔ مدل پیشنهادی با مدل‌های پایه، دو حالت مدل پیشنهادی، جمله‌بنیان و بافت‌بنیان با یکدیگر نیز مقایسه شد. نتایج آزمایش‌ها نشان داد که حالت بافت‌بنیان الگوریتم جداگرایانه در دو شاخص V و F کارایی بالاتری از حالت جمله‌بنیان را به‌دست آورده است که بر اساس آزمون تی دو جهته، تفاوت بین این دو حالت برای شاخص V از نظر آماری معنادار است. به‌دست آوردن بهترین نتیجه در حالت بافت‌بنیان به این مفهوم است که برای تعیین معنای واژه‌های هدف در زبان فارسی، بافت جایگاهی و واژه‌های مجاورِ واژهٔ هدف کفایت می‌کند. برای ارزیابی درونی خوشه‌بندی، از معیار سنجش تراکم و تیرگی داده استفاده شد و میانگین مقدار تراکم داده در تعداد خوشه‌های انتخاب‌شدهٔ واژه‌های هدفِ فارسی در دو حالت جمله‌بنیان و بافت‌بنیان با یکدیگر مقایسه شد. بر اساس نتایج، حالت جمله‌بنیان تراکم بالاتری نسبت به حالت بافت‌بنیان به‌دست آورد که تفاوت این دو مدل از نظر آماری معنادار است.

با مقایسهٔ ارزیابی درونی و برونی، دو نتیجهٔ متفاوت به‌دست آمد به این صورت که در ارزیابی درونی، مدل جمله‌بنیان و در ارزیابی برونی، مدل بافت‌بنیان کارایی بالاتری را نسبت به مدل دیگر به‌دست آورده است. ابتدا باید به خاطر داشت که ماهیت معیارهای ارزیابی با یکدیگر متفاوت است و به خاطر همین تفاوت است که معیارهای ارزیابی متفاوت معرفی می‌شود تا بتوان از جنبه‌های مختلف یک مسئله مورد بررسی قرار گیرد. این پژوهش نمونهٔ بارز تفاوت در نتیجه‌گیری است. دلیلی که می‌توان برای تفاوت نتیجهٔ ارزیابی‌ها در دو مدل پیشنهادی ذکر کرد، این است که در ارزیابی درونی، کیفیت خود الگوریتم خوشه‌بندی به‌صورت مستقل مورد ارزیابی قرار می‌گیرد؛ در حالی که در ارزیابی برونی، خروجی الگوریتم با یک دادهٔ استاندارد طلایی مقایسه می‌شود. بنابراین، درست است که تراکم و تیرگی خوشه‌های به‌دست‌آمده در مدل جمله‌بنیان بالاست، ولی ممکن است نتیجهٔ حاصل از تراکم بالای خوشه‌ها در مقایسه با دادهٔ استاندارد طلایی همسو نباشد. همچنین، این احتمال وجود دارد که وجود اطلاعات زیاد واژگانی جملات مدل جمله‌بنیان، به وجود نوفه[1] در داده منجر گردد و به گمراهی الگوریتم خوشه‌بندی

1. noise

بیانجامد. بنابراین، هرقدر از نوفه داده کاسته شود، به نتیجهٔ واقع‌گرایانه‌تر نزدیک می‌شویم.

نتیجهٔ نهایی به‌دست‌آمده از این پژوهش این است که می‌توان معنای واژهٔ هدف در بافت زبانی را با استفاده از روش‌های استنتاج استقرایی به‌دست آورد. انتخاب مدل پردازش داده، در سطح جمله یا بافت جایگاهی، با توجه به شاخص‌های ارزیابی متفاوت ممکن است به نتایج مختلف منتج شود. بنابراین، مدل پردازشی باید با توجه به اهمیت ویژگی‌های مورد نظر در سؤال پژوهش انتخاب گردد.

فهرست منابع

بی‌جن‌خان، محمود. ۱۳۸۳. نقش پیکرهٔ زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای. *مجله زبان‌شناسی* ۱۹ (۲): ۴۸-۶۷.

_____، و الهام علایی ابوذر. ۱۳۹۲. عمق خط فارسی. *پژوهش‌های زبانی* ۴ (۱): ۱-۱۹.

بی‌جن‌خان، محمود، و شهروز مرادزاده. ۱۳۸۳. هم‌نگاره‌های خط فارسی. در *مجموعه سخنرانی‌ها، گزارش‌ها و چکیده طرح‌های اولین کارگاه پژوهشی زبان فارسی و رایانه*. دانشگاه تهران، ۵۳-۶۳.

خسروی‌زاده،پروانه، و علی فارسی‌نژاد. ۱۳۹۱. ابهام‌زدایی از معنای کلمه با الگوریتم لسک ساده و گسترش‌یافته. در *مجموعه‌مقالات دومین هم‌اندیشی زبانشناسی رایانشی*. پژوهشگاه علوم انسانی و مطالعات فرهنگی، ۵۹-۷۸.

ذوالفقارکندری، زهره و طیبه موسوی میانگاه. ۱۳۹۴. ابهام‌زدایی واژگانی صفات مبهم در ترجمهٔ ماشینی: بررسی پیکره بنیاد. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۰ (۳): ۷۱۹-۷۳۵.

سلطانی، محمود، و هشام فیلی. ۱۳۸۷. استفاده از تکنیک ابهام‌زدایی معنایی واژگان در بازیابی بین‌زبانی اطلاعات. در *مجموعه‌مقالات چهاردهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران*. دانشگاه امیرکبیر.

شول، حمیدرضا، و رضا نورمندی‌پور. ۱۳۹۳. ارائه یک سیستم ابهام‌زدایی خودکار معنایی کلمات بر اساس سیستم دفاعی بدن انسان. در *مجموعه‌مقالات دومین همایش ملی پژوهش‌های کاربردی در علوم کامپیوتر و فناوری اطلاعات*. دانشکده مدیریت دانشگاه تهران.

قیومی، مسعود. ۱۳۹۸. ارائه یک روش مبتنی‌بر مدل زبانی برای واحدسازی پیکرهٔ فارسی. *زبان و زبان‌شناسی* ۱۴ (۲۷): ۲۱-۵۰.

کیانی‌نژاد، سپیده، حسین شیرازی، و سعیده سادات سدیدپور. ۱۳۹۵. تأثیر الگوریتم ابهام‌زدایی معنایی کلمه در دسته‌بندی سنجمان. در *اولین مسابقه کنفرانس بین‌المللی جامع علوم مهندسی در ایران*. بندر انزلی.

مسعودی، بابک، و سعید راحتی. ۱۳۹۴. رفع ابهام واژگان مبهم فارسی با مدل موضوعی LDA. *پردازش علائم هوشمند*، ۱۲ (۴): ۱۱۷-۱۲۵.

References

AleAhmad, A., H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian. 2009. Hamshahri: A standard Persian text collection. *Knowledge-Based Systems* 22 (5): 382–387.

Assi, S. M. 1997. Farsi linguistic database (FLDB). *International Journal of Lexicography* 10 (3): 5.

Bijankhan, M., J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi. 2011. Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*0 45 (2): 143–164.

Blei, D. M., T. L. Griffiths, M. I. Jordan, M. I. Jordan, and J. B. Tenenbaum. 2003a. Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems*. Whistler, British Columbia, Canada, MIT Press, pp: 17–24.

Blei, D. M., A. Ng, and M. Jordan. 2003b. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:1022–993 .

de Saussure, F. 1916. *Cours de linguistiquegénérale*. Lausanne, Paris: Payot.

Fakhrahmad S., A. Rezapour, M. ZolghadriJahromi, and M. Sadreddini.2011 . A new word sense disambiguation system based on deduction. In *Proceedings of the World Congress on Engineering*. London.

Fakhrahmad S., A. Rezapour, M. ZolghadriJahromi, and M. Sadreddini. 2012. A new fuzzy rule-based classification system for word sense disambiguation. *Intelligent Data Analysis* 16 (4): 633-648.

Fakhrahmad S., M. Sadreddini, and M. ZolghadriJahromi. 2014. A proposed expert system for word sense disambiguation: deductive ambiguity resolution based on data mining and forward chaining. *Expert Systems* 32 (2): 178-191.

Firth, J. R. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis (special volume of the Philological Society)* p. 1–32. Oxford: Blackwell.

Ghayoomi, M., S. Momtazi, and M. Bijankhan. 2010. A study of corpus development for Persian. *International Journal on Asian Language Processing* 20 (1): 17–33.

Hamidi, M. and A. Borji and S. ShiryGhidary. 2007. Persian word sense disambiguation. In *Proceedings of IEEE 15th Iranian Conference on Electrical Engineering*. Tehran. pp. 114-118.

Harris, Z. S. 1954. Distributional structure. *Word* 23 (10): 146–162.

Huang, E., R. Socher, C. D. Manning, and A. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the ACL*, volume 1, pp: 873–882. Jeju Island, Korea.

Jurafsky, D. and J. H. Martin. 2018. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (accessed July 2019).

Lau, Jey H., Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection, In *Proceedings of the 13the Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA: ACL, pp: 591-601.

Li, J. and D. Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the Conference on EMNLP*, pp: 1722–1732. Lisbon, Portugal.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics*, pp: 768–774. Montreal, Quebec, Canada.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp: 281–297.

Mahmoodvand, M., and M. Hourali. 2015. Persian word sense disambiguation corpus extraction based on word crawler method. In *International Journal of Advances in Computer Science* 4 (5): 101-106.

Manandhar, S., I. P. Klapaftis, D. Dligach, and S. S. Pradhan. 2010. SemEval-2010 task: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp: 63-68. Los Angeles, California.

MosaviMiangah, T., and A. DelavariKhalafi. 2005. Word sense disambiguation using target language corpus in a machine translation system. *Literary and Linguistic Computing* 20 (2): 237-249.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems* 26: 3111–3119.

Miller, G. A., and W. G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6 (1): 1–28.

Neelakantan, A., J. Shankar, A. Passos, and A. McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Processing of the Conference on EMNLP*, Doha, Qatar.

Pantel, P., and D. Lin. 2002. Discovering word senses from text. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, pp: 613–619. New York, USA.

Pennington, J., R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on EMNLP*, pp: 1532–1543. Doha, Qatar.

Rasekh, A. H., M. H. Sadreddini, and S. M. Fakhrahmad. 2014. Word sense disambiguation based on lexical and semantic features using Naïve Bayes classifier. *Journal of Computing and Security* 1 (2): 123-132.

Rekabsaz, N., S. Sabetghadam, M. Lupu, L. Andersson, and A. Hanbury. 2016. Standard test collection for English-Persian cross-lingual word sense disambiguation. In *Proceedings of the International Conference on Language Resource and Evaluation*, pp. 4176-4179. Portorož, Slovenia.

Rezapour, A., S. Fakhrahmad, M. Sadreddini and M. ZolghadriJahromi. 2014. An accurate word sense disambiguation system based on weighted lexical features. *Literary and Linguistic Computing* 29 (1): 74-88.

Riahi, N., and F. Sedghi. 2012. A Semi-supervised method for Persian homograph disambiguation. In *Proceedings of IEEE 20th Iranian Conference on Electrical Engineering*. pp. 748-751. Tehran.

Rosenberg, A. and J. Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the Joint Conference on EMNLP and CoNLL*, pp: 410–420. Prague, Czech Republic

Salton, G. M., A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18 (11): 613–620.

Shamsfard, M., A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi, E. Fekri, M. Monshizadeh, and S. M. Assi. 2010. Semi automatic development of Farsnet; The Persian wordnet. In *Proceedings of 5th Global WordNet Conference*. Mumbai, India.

Sarrafzadeh, B., N. Yakovets, N. Cercone, and A. An. 2011a. Cross-lingual word sense disambiguation for languages with scarce resources. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pp. 347-358. St. John's, Canada.

_____. 2011b. Towards automatic acquisition of a fully sense tagged corpus for Persian. In *Foundation of Intelligent Systems*, *International Symposium on Methodologies for Intelligent Systems*, pp. 449-455. Warsaw, Poland.

Shaoul, C., and C. Westbury. 2010. The Westbury Lab Wikipedia Corpus. http://www.psych.ualberta.ca/westburylab/downloads/westburylab.wikicorp.download.html (accessed July 2019).

Song, L., Z. Wang, H. Mi, and D. Gildea. 2016. Sense embedding learning for word sense induction. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, pp: 85–90. Berlin, Germany.

Van de Cruys, T. and M. Apidianaki. 2011. Latent semantic word sense induction and disambiguation. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*. 1476–1485. Portland, Oregon, USA.

Van de Cruys, T. 2010. Mining for Meaning: The Extraction of Lexico-semantic Knowledge from Text. PhD thesis, University of Groningen, The Netherlands.

Van Rijsbergen, C. J. 1979. *Information Retrieval*. Butterworth-Heinemann, Newton.

Wittgenstein, L. 1953. *Philosophical Investigations*. Blackwell Publishing Ltd, Oxford.

مسعود قیومی

متولد سال ۱۳۵۸ دارای مدرک تحصیلی دکتری در رشتهٔ زبان‌شناسی رایانشی از دانشگاه آزاد برلین، آلمان است. ایشان هم‌اکنون استادیار پژوهشکدهٔ زبان‌شناسی در پژوهشگاه علوم انسانی و مطالعات فرهنگی است. زبان‌شناسی رایانشی و پردازش زبان طبیعی، مدل‌سازی زبانی، یادگیری ماشینی، نحو و معناشناسی واژگانی از جمله علایق پژوهشی وی است.