دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

پایان‌نامه کارشناسی‌ارشد
علوم کامپیوتر - گرایش محاسبات نرم و هوش مصنوعی

تعیین نزدیک‌ترین کلمه به یک مفهوم فارسی به کمک فرهنگ واژگان وارونه

نگارش
آرمان ملک‌زاده لشکریانی

استادان راهنما
دکتر علی محدث خراسانی و دکتر امین غیبی

استاد مشاور
دکتر محمد بحرانی

شهریور ۱۳۹۹


بسم الله الرحمن الرحيم

صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تایید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

نکات مهم:

- نگارش پایان نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت پشت و رو(دورو) بلامانع است و انجام آن توصیه می شود.

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

به نام خدا

تعهدنامه اصالت اثر

تاریخ: شهریور ۱۳۹۹

اینجانب **آرمان ملک‌زاده لشکریانی** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

آرمان ملک‌زاده لشکریانی

امضا

تقدیم به پدر و مادر عزیزم که در تمامی مراحل زندگی یار و یاور من بوده‌اند. . .

سپاس‌گزاری

به ثمر رسیدن این پایان‌نامه مرهون حمایت‌های علمی و معنوی افرادی است که در مسیر انجام پژوهش و نگارش آن اینجانب را به طرق مختلف یاری نموده‌اند. لذا بر خود لازم می‌دانم که فرصت را غنیمت شمرده و مراتب قدردانی خود را از یکایک آن‌ها اعلام نمایم.

از آقایان دکتر علی محدث و دکتر امین غیبی کمال تشکر را دارم که به عنوان اساتید راهنما، همچون اعضای خانواده‌ای دلسوز همواره مرا از الطاف خود بهره‌مند ساخته‌اند. علاوه بر این دو عزیز، از آقای دکتر محمد بحرانی ممنونم که در مقام استاد مشاور، هیچگاه سوالات اینجانب را بی‌پاسخ نگذاشتند. همچنین از خانم دکتر مریم دانای طوس، دانشیار زبان‌شناسی دانشگاه گیلان سپاسگزارم که با انتقال تجربیات خود در زمینه زبان‌شناسی، اینجانب را در جهت انجام بهتر این پژوهش یاری نمودند. از دانشجویان رشته‌های ادبیات و زبان‌شناسی آن دانشگاه نیز تشکر می‌کنم که در ارزیابی روش‌های حل مسئله مورد بررسی در این پایان‌نامه با اینجانب همکاری نمودند. در پایان، قدردان پدر و مادرم هستم که حمایت‌های معنوی‌شان، همواره باعث دلگرمی اینجانب بوده است.

آرمان ملک‌زاده لشکریانی

شهریور ۱۳۹۹

چکیده

دسترسی به کلمات مناسب برای انتقال مفاهیم (دسترسی واژگانی) از ضروریات ایجاد ارتباط موثر میان انسان‌ها است. در راستای شبیه‌سازی کامپیوتری دسترسی واژگانی، می‌توان سیستمی را در نظر گرفت که امکان جستجوی کلمات مرتبط به یک مفهوم مشخص را برای کاربر فراهم می‌کند. این سیستم اصطلاحا فرهنگ واژگان وارونه نامیده می‌شود. در این پایان‌نامه، با استفاده از شبکه‌های عصبی مصنوعی، مدل‌های مختلفی برای پیاده‌سازی یک فرهنگ واژگان وارونه هوشمند مخصوص زبان فارسی ارائه می‌شود. این مدل‌ها باید توانایی نگاشت یک عبارت توصیفی به یک کلمه متناظر آن را داشته باشند. از همین رو، برای آموزش آن‌ها داده‌هایی به شکل «عبارت و کلمه» را از سه منبع زبانی فارسی استخراج می‌کنیم: لغت‌نامه‌ها (عمید، دهخدا و معین)، ویکی‌پدیا و شبکه واژگانی. هر مدل باید قادر باشد همه عبارات توصیفی متناظر یک کلمه را که از منابع مختلف استخراج شده‌اند، به آن نگاشت کند. همچنین، از آنجایی که لغت‌نامه‌های معین و عمید حداقل سی سال پس از لغت‌نامه دهخدا منتشر شده‌اند، تفاوت‌هایی از نظر انتخاب کلمات و ساختار جملات میان این لغت‌نامه‌ها وجود دارد. این تفاوت‌ها نیز باید توسط مدل مدیریت گردد. برای ارزیابی عملکرد سیستم پیشنهادی، بخشی از داده‌های استخراج‌شده از لغت‌نامه‌ها را انتخاب می‌کنیم. مدخل هر کلمه در لغت‌نامه شامل عباراتی در وصف آن کلمه است. لذا می‌توان کلمه متناظر هر مدخل را به مثابه نزدیک‌ترین کلمه به آن عبارات توصیفی بر مبنای قضاوت پدیدآورنده لغت‌نامه دانست. آزمایش‌های ما نشان می‌دهند که با انتخاب یک بازنمایی برداری مناسب برای کلمات و استفاده از حافظه کوتاه-مدت ماندگار به همراه لایه‌های توجه و تماما متصل، می‌توان مدلی را طراحی نمود که به خوبی یک فرهنگ واژگان وارونه را پیاده‌سازی می‌نماید. به طور دقیق‌تر، کلمه خروجی مدل به ازای دریافت عبارات توصیفی، از نظر قدرت انتقال مفاهیم موجود در عبارت ورودی، قابل قیاس با کلمه مندرج در مدخل لغت‌نامه و گاهی بهتر از آن می‌باشد.

واژه‌های کلیدی:

فرهنگ واژگان وارونه، جستجوی مفهوم، دسترسی واژگانی، پردازش زبان طبیعی، شبکه‌های عصبی مصنوعی

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فهرست جداول

جدول صفحه

فهرست نمادها

نماد	مفهوم
$\mathbb{R}^n$	فضای اقلیدسی با بعد n
e	عدد اویلر
$tanh$	تابع تانژانت هایپربولیک
$\odot$	ضرب مولفه به مولفه
UNK	نماد یک واحد ناشناخته در یک دنباله از کلمات
PAD	نماد یک واحد اضافه‌شده برای تنظیم طول یک دنباله از کلمات
cos	شباهت کسینوسی
$Acc_{exact}@k$	دقت در k بر حسب لیست کلمات صحیح
$Acc_{syn}@k$	دقت در k بر حسب لیست کلمات مترادف
FT	مدل بازنمایی برداری کلمات فارسی بر اساس معماری FastText
HAM-W2V	مدل بازنمایی کلمات فارسی بر اساس معماری Word2Vec و پیکره همشهری
WIKI-W2V	مدل بازنمایی کلمات فارسی بر اساس معماری Word2Vec و پیکره ویکی‌پدیای فارسی

فصل اول

مقدمه

گاهی اوقات انسان از انتخاب واژگان مناسب برای انتقال مفاهیم موردنظر خود باز می‌ماند. در این گونه مواقع، اغلب با تفصیل در کلام، بیان غیر مستقیم و یا انتخاب کلمات نزدیک اما متفاوت از نظر مفهومی، سعی می‌کند منظور خود را برساند. فرهنگ واژگان وارونه، سیستمی کامپیوتری است که در راستای کمک به بیان دقیق‌تر مفهوم موردنظر، به انسان کمک می‌کند با توصیف آنچه در ذهن دارد، مرتبط‌ترین کلمات به آن را بیابد.

در سال‌های اخیر با استفاده از شبکه‌های عصبی پیشرفت‌های زیادی در زمینه پردازش زبان‌های طبیعی حاصل شده است. تکنیک‌های پردازش زبان طبیعی مدرن، از بازنمایی برداری کلمات [۱] برای حل مسائل مختلف بهره می‌گیرند. آزمایش‌ها نشان داده است این گونه بازنمایی، روابط میان کلمات مختلف را به خوبی منعکس می‌نماید [۲۹]. پژوهشگران به کمک حافظه کوتاه-مدت ماندگار [۲] و شبکه‌های عصبی پیچشی [۳]، از این بازنمایی‌های برداری برای نگاشت دنباله‌ای از کلمات به دنباله‌ای دیگر (نگاشت دنباله-به-دنباله) و دسته‌بندی جملات استفاده نمودند [۴۷، ۲۳]. ترجمه متون یک زبان به زبانی دیگر، یکی از کاربردهای نگاشت دنباله-به-دنباله است. پژوهشگران با مشاهده نحوه ترجمه متون توسط انسان، دریافتند که برای تولید هر کلمه از دنباله خروجی، به بخش‌هایی از دنباله ورودی توجه بیشتری می‌شود. آن‌ها برای پیاده‌سازی این روند به صورت هوشمند، مکانیزم توجه [۴] را ارائه نمودند [۲]. گروه دیگری از پژوهشگران با استفاده از مکانیزم توجه، معماری «تبدیل‌کننده» [۵] را ابداع نمودند. آن‌ها با استفاده از این معماری جدید، موفق شدند ترجمه ماشینی هوشمند را با دقت بیشتری نسبت به گذشته و بدون استفاده از شبکه‌های عصبی پیچشی یا حافظه کوتاه-مدت ماندگار انجام دهند [۵۰]. اخیرا پژوهشگران مدل «بازنمایی‌های رمزگذار دوسویه بر پایه تبدیل‌کننده» [۶] را ارائه داده‌اند. این مدل می‌تواند به عنوان یک نقطه شروع برای هر مسئله در حوزه پردازش زبان طبیعی از جمله پرسش و پاسخ [۷] به کار گرفته شود [۱۱].

در این پایان‌نامه، با استفاده از تکنیک‌های پردازش زبان طبیعی مدرن سعی داریم نزدیک‌ترین کلمه به یک مفهوم در زبان فارسی را تعیین کنیم. برای این کار، از یک فرهنگ واژگان وارونه کمک خواهیم گرفت. این فرهنگ واژگان وارونه، دنباله‌ای از کلمات را دریافت می‌کند و در ازای آن یک کلمه را تحویل

[۱] Word vector representations
[۲] Long short-term memory
[۳] Convolutional neural networks
[۴] Attention mechanism
[۵] Transformer
[۶] Bidirectional encoder representations from transformer
[۷] Question answering

می‌دهد. از این رو می‌توان مسئله مذکور را نوعی نگاشت دنباله-به-دنباله دانست که در آن دنباله خروجی، تنها متشکل از یک کلمه است.

در ادامه، صورت مسئله به شکل دقیق بیان خواهد شد. سپس کاربردهای آن ذکر می‌گردد و در نهایت ساختار فصل‌های بعدی پایان‌نامه بیان خواهد شد.

۱-۱ صورت مسئله

۱-۱-۱ تعیین نزدیک‌ترین کلمه به یک مفهوم فارسی

فرض کنید $V = \{w_۱, w_۲, \ldots, w_n\}$ مجموعه کلمات موجود در یک زبان باشد. یک دنباله از این کلمات مانند $S = (s_۱, s_۲, \ldots, s_m)$ به عنوان ورودی داده می‌شود. هدف، یافتن کلمه‌ای مانند w_t است که بیشترین ارتباط معنایی را به دنباله داده‌شده داشته باشد. اگر این دنباله از کلمات یک مفهوم در زبان فارسی را توصیف کنند، کلمه w_t را نزدیک‌ترین کلمه به آن مفهوم فارسی می‌نامیم.

۱-۱-۲ فرهنگ واژگان وارونه

فرهنگ واژگان (لغت‌نامه) یک منبع زبانی متشکل از اطلاعات مربوط به کلمات یک زبان انسانی است و برای هر کلمه، مواردی از قبیل نحوه تلفظ، معنی کلمه [۸] و گاهی کاربردهای آن را در بر دارد. معمولا در یک فرهنگ واژگان کلمات بر اساس ترتیب حروف الفبا مرتب شده‌اند. از این نوع فرهنگ واژگان بیشتر برای یادگیری معانی کلمات استفاده می‌شود؛ اما فرهنگ واژگان انواع دیگری نیز دارد.

فرهنگ واژگان وارونه یک نوع خاص از فرهنگ واژگان است که با اهداف متفاوتی پدید می‌آید. در نسخه‌های مکتوب این نوع فرهنگ واژگان، کلمات بر اساس ترتیب حروف الفبا و از حرف آخر به حرف اول مرتب شده‌اند. به همین دلیل، گاهی به این نسخه‌ها فرهنگ زانسو (از آن سو) گفته می‌شود. لغت فُرس نوشته اسدی طوسی [۵۵] و فرهنگ فارسی زانسو نوشته کشانی [۵۶] دو نمونه فرهنگ زانسو برای زبان فارسی هستند.

در دهه‌های اخیر، نسخه‌های دیجیتال فرهنگ واژگان وارونه برای زبان‌هایی از جمله انگلیسی [۵۲] ، ژاپنی [۵] ، ترکی [۱۳] و فرانسوی [۱۲] ارائه شده است. این نسخه‌های دیجیتال، سیستم‌های کامپیوتری هستند که با دریافت یک عبارت توصیفی از کاربر، نزدیک‌ترین کلمه به آن عبارت را تعیین

[۸] Word sense

می‌کنند. ساخت این فرهنگ واژگان‌های وارونه، در ابتدا با استفاده گراف‌های دانش و یا سیستمی شامل کلیه اطلاعات مربوط به کلمات یک زبان صورت می‌گرفت. در سال‌های اخیر و با توجه به پیشرفت‌های حاصل‌شده با استفاده از شبکه‌های عصبی در زمینه پردازش زبان طبیعی، فرهنگ واژگان وارونه به شکل یک سیستم هوشمند طراحی شده است. با وجود این دستاوردها، زبان فارسی از داشتن نسخه دیجیتال این منبع زبانی محروم مانده است. در این پایان‌نامه، روند پیاده‌سازی یک فرهنگ واژگان وارونه هوشمند برای زبان فارسی شرح داده می‌شود. کاربر می‌تواند مفهومی را که در ذهن دارد، به وسیله یک عبارت توصیف نماید. فرهنگ واژگان وارونه هوشمند، با پردازش این عبارت توصیفی، نزدیک‌ترین کلمه به آن مفهوم را به کاربر ارائه می‌کند.

۱-۲ کاربردها

تعیین نزدیک‌ترین کلمه به یک مفهوم به کمک فرهنگ واژگان وارونه، کاربردهای متفاوتی دارد. برخی از این کاربردها در ادامه ذکر می‌شوند.

- **انتخاب کلمات کلیدی برای استفاده توسط موتورهای جستجو**

یک موتور جستجوی وب، دنباله‌ای از کلمات را از کاربر دریافت می‌کند و در پاسخ، مرتبط‌ترین صفحات وب را بازمی‌گرداند. هر چه این کلمات حاوی مفاهیم دقیق‌تری باشند، موتور جستجو بهتر عمل می‌نماید. در این مواقع اصطلاحا این کلمات را «کلمات کلیدی» می‌نامیم. تشخیص کلیدی‌بودن یک کلمه، برای کاربر غیر متخصص چالش‌برانگیز است. این کاربر می‌تواند هر آنچه را که در ذهن خود از مفهوم موردنظر متصور است، به شکل دنباله‌ای از کلمات به یک فرهنگ واژگان وارونه بدهد و در ازای آن، لیستی از کلمات کلیدی مرتبط را دریافت نماید. سپس با استفاده از این کلمات، دنباله کلمات اولیه را تغییر دهد. بدین ترتیب، احتمال یافتن صفحات مرتبط توسط موتور جستجو افزایش می‌یابد.

- **کمک به نویسندگان در انتخاب کلمات**

یکی از اعمال مورد ستایش در نویسندگی، پرهیز از اطناب و درازگویی است. این امر از گذشته‌های دور مورد توجه شعرا و نویسندگان بوده است. آن چنان که حافظ در یکی از غزلیات خود می‌فرماید:

«بیا و حال اهل درد بشنو / به لفظ اندک و معنی بسیار»

این بیت نشان می‌دهد که از گذشته تلاش نویسندگان بر آن بوده است که معانی و مفاهیم در غالب عبارات کوتاه گنجانده شوند. اما گاهی ممکن است آن‌ها در به یاد آوردن کلمات متناظر مفاهیم موردنظر خود دچار مشکل شوند. در این صورت می‌توانند آن مفاهیم را به صورت عبارات توصیفی بیان کنند و با استفاده از فرهنگ واژگان وارونه، به کلمات متناظر آن‌ها دست یابند.

- **کمک به زبان‌آموزان**

برای یادگیری یک زبان جدید، معمولا زبان‌آموزان از کلمات ساده شروع می‌کنند. غالبا این کلمات حاوی معانی پیچیده نیستند. اما با ترکیب آن‌ها می‌توان یک مفهوم پیچیده‌تر را توصیف نمود. فرض کنید دنباله s از ترکیب چند کلمه ساده تشکیل شده باشد و مفهوم c را توصیف کند. اگر یک فرهنگ واژگان وارونه s را دریافت کند، کلمه‌ای مانند w را تحویل می‌دهد که مفهوم c را در بر دارد. بدین ترتیب، زبان‌آموز می‌تواند کلمه w را نیز فرا بگیرد. این روند به خصوص برای کسانی که به تازگی به یک کشور مهاجرت کرده‌اند، کاربرد دارد. آن‌ها می‌توانند به کمک یک فرهنگ واژگان وارونه، مفاهیم مورد نظر خود را به سادگی در ارتباط با بومیان به کار گیرند.

- **افزایش توان‌مندی و هوشمندسازی یک چت‌بات**

یکی از راه‌های ارائه خدمات غیرحضوری به مشتریان هر شرکت یا سازمان، پشتیبانی برخط می‌باشد. در این نوع ارائه خدمت، مشتری به تارنمای وب معرفی‌شده توسط سرویس‌دهنده مراجعه می‌کند و با کارشناسان در قالب یک گفتگوی متنی، تعامل می‌نماید. فرض کنید کاربری از این خدمات استفاده می‌کند و پیام متنی زیر را برای یک شرکت اپراتور تلفن همراه ارسال می‌نماید:

«با من تماس می‌گیرند. اما من نمی‌توانم با دیگران تماس بگیرم»

در حالت عادی، یک کارشناس این پیام را می‌خواند و ممکن است به او بگوید که اعتبار سیم‌کارت او به پایان رسیده است. با ارائه داده‌های مناسب و آموزش یک فرهنگ واژگان وارونه، می‌توان کلمات «اتمام» و «شارژ» را از این سیستم به عنوان خروجی متناظر پیام متنی کاربر دریافت نمود. سپس یک سیستم دیگر می‌تواند با استفاده از این کلمات، جمله کاملی مانند «شارژ شما به اتمام رسیده است» را تولید نماید. با این کار، دیگر نیازی به استخدام کارشناس وجود نخواهد داشت و در هزینه‌های شرکت یا سازمان سرویس‌دهنده صرفه‌جویی خواهد شد.

- **حل مسئله نوک زبان**

گاهی اوقات انسان کلمه‌ای را به خاطر نمی‌آورد. اما کلماتی مشابه از نظر معنی و یا فرم نوشتاری به ذهن او می‌رسند. این مسئله در حوزه زبان‌شناسی به عنوان «پدیده نوک زبان» شناخته می‌شود [۷]. برای حل این مسئله می‌توان از یک فرهنگ واژگان وارونه استفاده کرد. شخص کلماتی را که به خاطر می‌آورد به عنوان ورودی به فرهنگ واژگان وارونه می‌دهد و در پاسخ، کلمه‌ای را که به خاطر نمی‌آورده تحویل می‌گیرد.

- **حل هوشمندانه جدول‌ها**

حل یک جدول عبارت است از پرکردن خانه‌های آن با توجه به توضیحاتی که در کنار جدول درج می‌شود. این توضیحات همواره از یک یا چند کلمه مانند $w_1, w_2, \ldots, w_k$ تشکیل شده‌اند که یک کلمه دیگر مانند w' را توصیف می‌کنند. تعداد کاراکترهای کلمه w' نیز بر اساس خانه‌های جدول مشخص شده است. فرض کنید خانه‌های $c_1, c_2, \ldots, c_m$ به کلمه w' اختصاص داده شده باشند. اگر دنباله $< w_1, w_2, \ldots, w_k >$ به عنوان ورودی به یک فرهنگ واژگان وارونه داده شود و از خروجی آن تنها کلمات به طول m استخراج گردد، لیست کلماتی که می‌توانند خانه‌های $c_1, c_2, \ldots, c_m$ را پر کنند مشخص می‌شود. بدین ترتیب می‌توان با یافتن یک مقداردهی به خانه‌های جدول که مطابق با تمامی توضیحات آن خانه‌ها باشد، جدول را حل نمود.

۱-۳ چالش‌ها

به واسطه عوامل مختلف، چالش‌هایی در مسیر طراحی فرهنگ واژگان وارونه هوشمند برای زبان فارسی وجود دارد. در این بخش، این چالش‌ها را بر اساس منشا ایجادشان دسته‌بندی کرده و شرح می‌دهیم. در روند حل مسئله، این چالش‌ها را مد نظر قرار خواهیم داد.

۱-۳-۱ داده‌ها

در این پژوهش، از داده‌های موجود در لغت‌نامه‌های فارسی به عنوان منبع اصلی استخراج نمونه‌هایی برای آموزش مدل‌های مختلف بهره خواهیم برد. تعارض میان معانی مختلف کلمات در لغت‌نامه‌ها و تغییرات زبان در طول زمان، دشواری‌هایی را برای حل مسئله ایجاد می‌کند. در بخش ۲-۳ به طور مفصل به این موارد پرداخته شده است.

۱-۳-۲ طراحی مدل

با توجه به اینکه در یک عبارت توصیفی کلمات دارای ترتیب هستند، مدل‌های پیشنهادی برای حل مسئله باید این ترتیب را در نحوه پردازش اطلاعات لحاظ کنند. همچنین، خروجی مدل‌ها باید کلمات یا اشیائی باشند که به نوعی آن‌ها را نمایش دهند.

۱-۴ نوآوری‌ها

در این پایان‌نامه، استفاده از شبکه‌های عصبی برای طراحی یک فرهنگ واژگان وارونه فارسی مورد بررسی قرار می‌گیرد. خلاصه نوآوری‌های این پژوهش بدین شرح است.

- استخراج یک مجموعه داده جدید حاوی عبارات توصیفی متناظر کلمات زبان فارسی (فصل ۳)
- ارائه یک مجموعه داده حاوی مجموعه کلمات مترادف برای زبان فارسی (فصل ۳)
- پیشنهاد مدلی بر پایه حافظه کوتاه-مدت ماندگار و لایه توجه برای تعیین نزدیک‌ترین کلمه به یک مفهوم فارسی (فصل ۴)

۱-۵ ساختار فصل‌ها

در این بخش، ساختار فصل‌های باقیمانده این پایان‌نامه شرح داده می‌شود. در فصل ۲ ، پژوهش‌های مرتبط در زمینه طراحی یک فرهنگ واژگان وارونه شرح داده می‌شود. فصل ۳ به بررسی داده‌های موردنیاز برای این پژوهش به همراه نحوه جمع‌آوری و آماده‌سازی آن‌ها می‌پردازد. در فصل ۴ ، معماری مدل‌های پیشنهادی برای حل مسئله تبیین می‌گردد. فصل ۵ به آزمایش‌های صورت گرفته روی داده‌ها با توجه به مدل‌های پیشنهادی اختصاص داده شده است. در نهایت در فصل ۶ به مهم‌ترین نتایج بدست‌آمده از این پژوهش اشاره می‌شود و پیشنهاداتی برای بهبود عملکرد سیستم ذکر می‌گردد.

فصل دوم

پیشینه پژوهش

در طول دو دهه اخیر، روش‌های متفاوتی برای تعیین نزدیک‌ترین کلمه (یا کلمات) به یک مفهوم پیشنهاد شده است. به طور کلی در این روش‌ها با سه رویکرد به حل مسئله پرداخته شده است. در رویکرد اول، ابتدا یک گراف دانش تشکیل می‌شود و یا از یک گراف دانش از قبل ساخته‌شده استفاده می‌گردد. در این رویکرد، معمولا کلمات به عنوان رئوس گراف در نظر گرفته می‌شوند و طول مسیر میان آن‌ها محاسبه می‌شود. در رویکرد دوم، از سیستم‌هایی استفاده می‌شود که یک پیکره [۱] نسبتا بزرگ در آن‌ها ذخیره شده است و برای هر عبارت توصیفی که از کاربر دریافت شده باشد، با معیارهای آماری بر مبنای آن پیکره، میزان شباهت آن به کلمات زبان محاسبه می‌گردد. رویکرد سوم که در سال‌های اخیر بیشتر مورد استفاده قرار گرفته است، با استفاده از شبکه‌های عصبی به حل مسئله می‌پردازد. در این فصل، پژوهش‌های صورت گرفته با هر یک از این رویکردها مرور می‌شود.

۲-۱ استفاده از یک گراف دانش

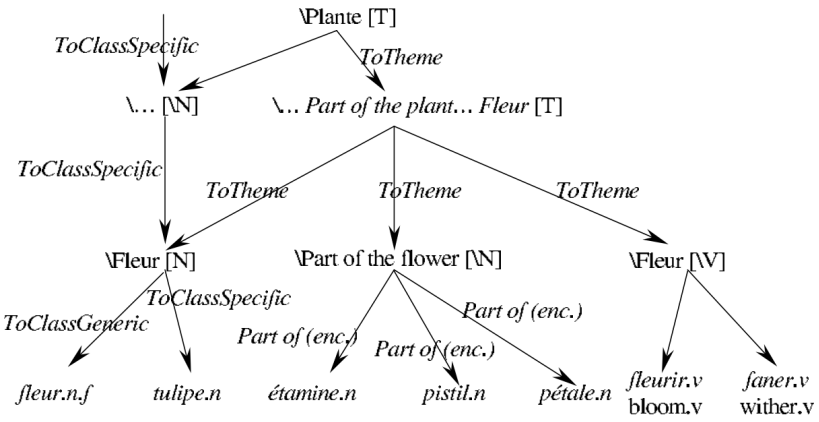
Dutoit و Nugues [۱۲] پیشنهاد دادند برای ساخت یک فرهنگ واژگان وارونه مخصوص زبان فرانسوی، از یک گراف دانش شامل مفاهیم و کلمات آن زبان استفاده شود. آن‌ها برای ایجاد این گراف، از یک منبع زبانی به نام «لغت‌نامه کامل» [۲] استفاده کردند. در لغت‌نامه کامل، کلمات یک زبان بر اساس مفاهیمی که در بر دارند، دسته‌بندی شده و در قالب یک گراف جهت‌دار از نوع درخت نمایش داده شده‌اند که برگ‌های آن کلمات هستند و مفاهیم، بقیه گره‌های آن را تشکیل می‌دهند. این دو پژوهشگر با گسترش این گراف، گرافی با ساختار مشابه بدست آوردند.

در ساختار درختی این گراف، مفاهیم کلی‌تر، در ارتفاع بالاتری قرار گرفته‌اند و در پایین‌ترین ارتفاع، کلمات قرار دارند. فرض کنید یال (u, v) در این گراف وجود داشته باشد. گوییم v فرزند u و u والد v است. همچنین، گوییم w یکی از نیاکان درجه k از v است؛ اگر کوتاه‌ترین مسیر از w به v شامل k راس به جز w و v باشد. برای تعیین فاصله میان دو کلمه (برگ) در این گراف، این دو پژوهشگر از میان نیاکان مشترک رئوس متناظر آن کلمات، راسی مانند t را که دارای کمترین درجه است، در نظر گرفتند. سپس با فرض بدون‌جهت بودن یال‌های درخت، بر اساس مسیری میان رئوس متناظر کلمات که از راس t عبور کند، معیاری از دوری و نزدیکی آن کلمات ارائه نمودند. برای تعمیم این معیار به نحوی که با استفاده از آن بتوان فاصله میان یک کلمه و یک عبارت را محاسبه نمود، آن پژوهشگران از روشی برای ادغام رئوس

[۱] Corpus

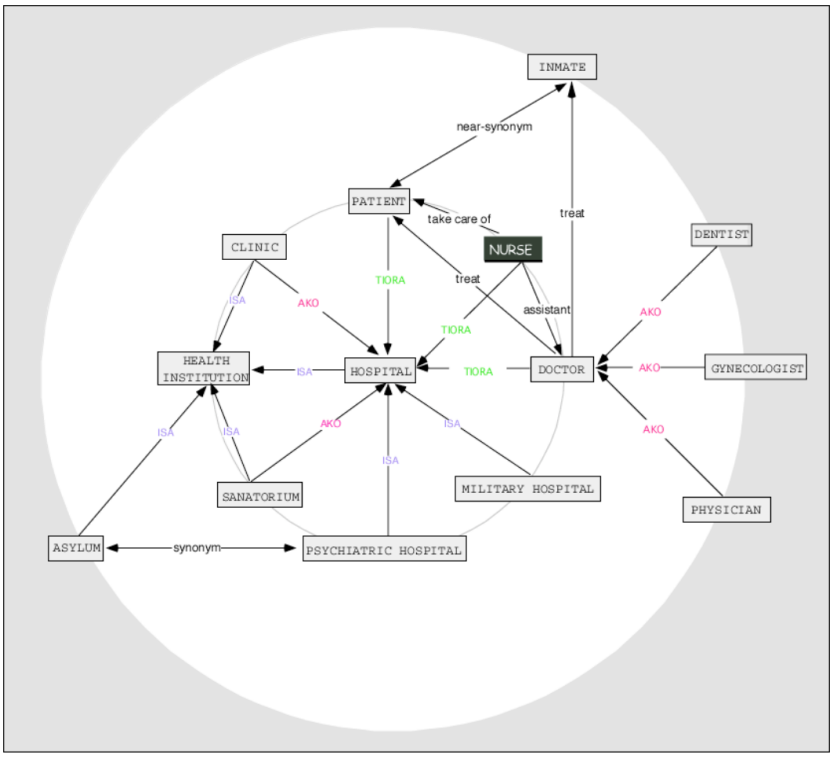
[۲] Integral dictionary

متناظر کلمات موجود در عبارت استفاده نمودند. بدین طریق با محاسبه فاصله هر کلمه موجود در زبان با یک عبارت داده‌شده که یک مفهوم را توصیف می‌کند، نزدیک‌ترین کلمات به آن مفهوم تعیین شد. شکل ۲-۱ بخشی از گراف لغت‌نامه کامل مربوط به کلمه fleur (گُل/ گُل‌کاشتن) را نشان می‌دهد.



شکل ۲-۱: بخشی از گراف لغت‌نامه کامل مورد استفاده Dutoit و Nugues [۱۲]

دو سال پس از طرح این ایده، Zock و Bilac [۵۲] پیشنهاد ساخت یک گراف دانش با ساختاری دیگر را دادند و ادعا نمودند پس از ساخت این گراف، کاربران می‌توانند با جستجوی کلمه به کلمه در آن، به کلمه نهایی مدنظر خود دست یابند؛ به اعتقاد آن‌ها، کاربر ممکن است کلمه‌ای را فراموش کرده باشد؛ اما کلمات دیگری را به خاطر می‌آورد که مرتبط با کلمه اصلی هستند. کلمات و مفاهیم در گراف پیشنهادی آن‌ها، گره‌ها را تشکیل می‌دهند و یال‌ها، ارتباط میان آن‌ها را مشخص می‌کنند. به طور مثال، از یک گره نماینده کلمه «دندان‌پزشک» به سوی گره دیگری که از کلمه «دکتر» نمایندگی می‌کند، یک یال وجود دارد که نماینده رابطه «نوعی از» است؛ زیرا «دندان‌پزشک»، نوعی «دکتر» می‌باشد. به عقیده این دو پژوهشگر، وجود رابطه میان کلمات مختلف می‌تواند با بررسی یک پیکره بزرگ به صورت خودکار مشخص شود. اما تعیین نوع رابطه، چالش‌برانگیز است و نیاز به یک منبع زبانی غنی دارد. شکل ۲-۲ بخشی از گرافی را نشان می‌دهد که Zock و Bilac به طور فرضی برای جستجوی کلمه Nurse (پرستار) در نظر گرفتند.



شکل ۲-۲: بخشی از گراف مفاهیم و کلمات مدنظر Zock و Bilac [۵۲]

گرچه Zock و Bilac ایده فوق را مطرح نمودند، اما پیاده‌سازی از آن ارائه نکردند. دو سال پس از طرح این ایده، Ferret با همکاری Zock [۱۶] و با استفاده از اخبار چاپ‌شده در یک روزنامه فرانسوی و یک ابزار تقسیم‌بندی موضوع [۳] به نام TOPICOLL [۱۵] ، گراف پیشنهادی آن‌ها را ایجاد نمود.

شبکه واژگانی یک پایگاه داده واژگانی برخط است که برای استفاده تحت کنترل یک برنامه کامپیوتری طراحی شده است و اسامی، افعال، صفات و قیدها را به مجموعه‌ای از مترادف‌ها پیوند می‌دهد که خود از طریق روابط معنایی پیوند دارند و تعاریف کلمات را معین می‌کنند [۳۲].

Méndez و همکاران [۲۸] با بهره‌گیری از شبکه واژگانی [۴] زبان انگلیسی، به هر کلمه از آن زبان یک بردار نسبت دادند. در شبکه واژگانی زبان انگلیسی، اسامی با یک ساختار سلسله مراتبی قرار گرفته‌اند. این پژوهشگران ۲۵ اسم که در کمترین ارتفاع قرار گرفته‌اند را به عنوان مبنایی برای محاسبه بردارها در نظر گرفته‌اند. برای هر کلمه موجود در زبان انگلیسی، فاصله آن با هر یک از این ۲۵ اسم با توجه به یکی از نیاکان مشترک آن‌ها در شبکه واژگانی که دارای کمترین درجه است، سنجیده می‌شود و به عنوان یک مولفه در بردار آن کلمه قرار می‌گیرد. بدین ترتیب، برای هر کلمه یک بردار محاسبه می‌شود.

[۳] Topic Segmentation
[۴] WordNet

اگر یک عبارت شامل کلماتی باشد، میانگین بردار آن کلمات به عنوان بردار متناظر آن عبارت شناخته می‌شود. با مقایسه بردار عبارت و بردار تک تک کلمات زبان، می‌توان نزدیک‌ترین کلمات به آن عبارت را مشخص نمود.

Thorat و Choudhari [۴۹] با پردازش عبارات توصیفی موجود در یک لغت‌نامه، یک گراف ایجاد کردند که رئوس آن کلمات زبان انگلیسی هستند. در لغت‌نامه، هر کلمه با یک عبارت توصیف شده است. فرض کنید p_w و $p_{w'}$ به ترتیب عباراتی در لغت‌نامه باشند که کلمات w و w' را توصیف می‌کنند. متناظر w و w' دو راس در گراف وجود دارد. اگر w در $p_{w'}$ وجود داشته باشد، یک یال راس متناظر w را به راس متناظر w' متصل می‌کند. از آنجایی که از برخی کلمات برای توصیف کلمات دیگر استفاده نمی‌شود، ممکن است گراف همبند نباشد. در این صورت، شرط جهت‌دار بودن تا حدودی نادیده گرفته می‌شود تا همبندی گراف تضمین گردد. طول کوتاه‌ترین مسیر میان رئوس متناظر دو کلمه، میزان شباهت آن‌ها را مشخص می‌کند. میزان شباهت یک کلمه با یک عبارت، بر اساس مجموع شباهت‌های آن کلمه با کلمات عبارت محاسبه می‌گردد.

Reyes-Magaña و همکاران [۴۰] با استفاده از «هنجارهای انجمنی» [۵] یک گراف از کلمات ساختند و سپس با استفاده از الگوریتم‌های Betweenness Centrality [۱۷] و PageRank [۳۴] میزان ارتباط کلمات عبارت ورودی با هر یک از کلمات زبان انگلیسی را محاسبه نمودند. هنجارهای انجمنی، کلماتی هستند که پس از شنیدن یک کلمه، به ذهن یک انسان بومی می‌رسد.

۲-۲ جستجو در یک پیکره

Bilac و همکاران [۵] با استفاده از یک فرهنگ مفاهیم زبان ژاپنی، روشی برای نمایش هر عبارت (دنباله از کلمات) به شکل یک بردار ارائه دادند. در روش آن‌ها، هر مولفه از آن بردار، متناظر یک کلمه از عبارت است و مقدار آن مولفه، بر اساس میزان رخداد آن کلمه در فرهنگ مفاهیم بدست می‌آید. فرهنگ مفاهیم زبان ژاپنی، شامل مفاهیم آن زبان و عبارات توصیفی متناظر آن‌ها می‌باشد. پژوهشگران مذکور با مقایسه بردار متناظر عبارت توصیفی داده‌شده توسط کاربر و بردار عبارات توصیفی موجود در فرهنگ مفاهیم، نزدیک‌ترین مفاهیم به مفهوم موردنظر کاربر را تعیین کردند.

El-Kahlout و Oflazer [۱۳] برای جستجو در یک لغت‌نامه زبان ترکی با فرض دریافت یک عبارت توصیفی از کاربر، روشی بسیار ساده بر اساس تطابق کلمات پیشنهاد کردند. آن‌ها تعداد کلمات مشترک

[۵] Association norms

میان عبارت توصیفی کاربر و هر عبارت توصیفی موجود در لغت‌نامه را به عنوان معیاری از شباهت این عبارات در نظر گرفتند. در یافتن این کلمات مشترک، این پژوهشگران به مجموعه ترادف هر کلمه نیز توجه داشتند و آن را از طریق شبکه واژگانی زبان ترکی بدست آوردند. همچنین، از روی هر عبارت توصیفی کاربر، با حذف تعدادی از کلمات آن، عبارات توصیفی دیگری نیز بدست آوردند و برای هر عبارت بدست‌آمده، روند پیشنهادی خود را تکرار کردند.

فرض کنید p عبارت توصیفی دریافت‌شده از کاربر باشد که حاوی n کلمه است. با حفظ ترتیب رخداد کلمات این عبارت و حذف یک یا تعدادی از آن‌ها، می‌توان $۲^n - ۱$ عبارت جدید بدست آورد. این عبارات از نظر میزان اطلاعات مفیدی که در بر دارند، با یکدیگر متفاوت هستند. برای محاسبه میزان مفیدبودن هر عبارت، پژوهشگران مذکور معیاری بر اساس میزان رخداد کلمات آن عبارت در کل پیکره لغت‌نامه و طول آن عبارت ارائه نمودند. بر این اساس، نتایج حاصل از هر عبارتی که حاوی اطلاعات مفیدتری باشد، زودتر به کاربر تحویل داده می‌شود.

Zock و Schwab [۵۳] از پیکره ویکی‌پدیای انگلیسی برای حل مسئله استفاده کردند. آن‌ها به صورت تصادفی متن ۱۰۰۰ صفحه ویکی‌پدیا را استخراج نمودند. فرض کنید کاربر عبارت توصیفی p شامل کلمات $w_۱, w_۲, \ldots, w_k$ را به عنوان ورودی در نظر گرفته باشد. همچنین، فرض کنید t_w یک کلمه به زبان انگلیسی باشد. در این روش، برای هر کلمه مانند t_w تعداد پاراگراف‌هایی از پیکره ویکی‌پدیا که شامل کلمات $t_w, w_۱, w_۲, \ldots, w_k$ هستند، به عنوان معیاری برای سنجش میزان نزدیکی عبارت p به کلمه t_w در نظر گرفته شده است.

Shaw و همکاران [۴۵] معیاری برای مقایسه عبارت توصیفی کاربر با هر یک از عبارات موجود در یک لغت‌نامه ارائه دادند. آن‌ها برای مقایسه دو عبارت، میزان شباهت هر یک از کلمات موجود در عبارت اول را با هر یک از کلمات موجود در عبارت دوم محاسبه نمودند. برای محاسبه شباهت میان دو کلمه، این پژوهشگران از محل قرارگیری آن کلمات در شبکه واژگانی استفاده نمودند. به طور خاص، برای آن کلمات، یکی از نیاکان مشترک آن‌ها که دارای کمترین درجه می‌باشد در نظر گرفته شد. همچنین، ارتفاع هر یک از کلمات در شبکه واژگانی مورد توجه قرار گرفت. لازم به ذکر است در روش این دو پژوهشگر، عبارت توصیفی کاربر قبل از آن که با عبارات موجود در لغت‌نامه مقایسه گردد، با توجه به روابط زیرشمولی [۶]، فراشمولی [۷] و ترادف [۸]، گسترش می‌یابد. پس از انجام مقایسه‌ها، اگر تعداد کلمات

[۶] Hyponymy
[۷] Hypernymy
[۸] Synonymy

خروجی سیستم از حد مشخصی بیشتر نباشد، با حذف بعضی کلمات از عبارت توصیفی کاربر، دامنه جستجو گسترش می‌یابد و جستجو تکرار می‌شود. این روند تا زمانی ادامه پیدا می‌کند که حداقل ۲ کلمه در عبارت توصیفی کاربر باقی بماند.

۲-۳ استفاده از شبکه‌های عصبی

Hill و همکاران [۲۱] با بهره‌گیری از شبکه‌های عصبی عمیق و نمایش کلمات به شکل بردار، مدل‌هایی را برای تعیین نزدیک‌ترین کلمات به یک مفهوم ارائه نمودند. در پژوهش آن‌ها فرض شده است که کاربر می‌تواند به وسیله یک عبارت، مفهوم موردنظر خود را توصیف نماید. برای هر کلمه از آن عبارت، یک بردار در نظر گرفته می‌شود. بر این اساس، آن‌ها دو معماری را برای حل مسئله پیشنهاد نمودند. در معماری اول، میانگین بردارهای کلمات، به عنوان بردار متناظر کل عبارت در نظر گرفته می‌شود. در حالی که در معماری دوم، بردار متناظر عبارت از طریق نوعی شبکه عصبی به نام حافظه کوتاه-مدت ماندگار بدست می‌آید [۲۲]. در هر دو معماری، بردار متناظر کل عبارت از طریق یک لایه تماما متصل [۹] به یک فضای دیگر تصویر [۱۰] می‌شود. لایه تماما متصل در واقع یک تابع غیرخطی را پیاده‌سازی می‌نماید. در نهایت می‌توان بردار هر کلمه را با بردار متناظر هر عبارت مقایسه نمود. بدین ترتیب، نزدیک‌ترین کلمات به آن عبارت (مفهوم) تعیین می‌گردد. این دو معماری، به عنوان معماری‌های پایه در این پژوهش مورد استفاده قرار گرفته‌اند.

Morinaga و Yamaguchi [۳۳] با آموزش یک شبکه عصبی پیچشی برای پیش‌بینی موضوع عبارت توصیفی کاربر و تلفیق آن با شبکه عصبی پیشنهادی Hill و همکاران [۲۱] ، عملکرد آن سیستم را بهبود بخشیدند. آن‌ها برای آموزش شبکه عصبی پیچشی، از داده‌های شبکه واژگانی زبان انگلیسی استفاده نمودند. در این شبکه واژگانی، برای هر کلمه، معانی مختلف ذکر شده است. برای هر معنی، یک عبارت توصیفی و همچنین موضوع آن موجود است. با کمک این داده‌ها، پژوهشگران مذکور توانستند سیستمی را طراحی کنند که ابتدا موضوعات مرتبط با عبارت توصیفی کاربر را شناسایی می‌نماید. بدین ترتیب، کلماتی که مرتبط با آن موضوعات نباشند، از لیست کلمات خروجی سیستم حذف می‌شوند.

پیله‌ور [۳۶] با استفاده از شبکه واژگانی زبان انگلیسی و تکنیک DeConf [۳۷] برای هر معنی از یک کلمه، برداری جدا بدست آورد. سپس با آموزش مدل ارائه‌شده توسط Hill و همکاران [۲۱] نشان

[۹] Fully connected layer

[۱۰] Projection

داد تفکیک معانی مختلف کلمه، باعث بهبود عملکرد سیستم می‌شود.

Hedderich و همکاران [۲۰] برای تفکیک معانی مختلف، از یک لایه توجه [۵۰] استفاده نمودند. در مدل پیشنهادی آن‌ها، لایه توجه بر اساس عبارت ورودی کاربر، برای هر کلمه یک بردار متناظر با معنی آن را انتخاب می‌نماید. در این پژوهش، از بردارهای ارائه‌شده توسط پیله‌ور [۳۶] استفاده شده است.

Zheng و همکاران [۵۱] در چند ماه اخیر با ترکیب چند شبکه عصبی، روشی جدید برای حل مسئله ارائه نمودند. به عقیده آن‌ها، کلمات یک عبارت دارای اهمیت متفاوتی هستند. برای تشخیص میزان اهمیت هر کلمه در یک عبارت، آن‌ها از یک لایه توجه استفاده نمودند. سپس با استفاده از یک حافظه کوتاه-مدت ماندگار دوسویه [۱۱] ، نمایشی برداری برای عبارت بدست آوردند. همچنین، با استفاده از چند شبکه پرسپترون تک لایه، جزء کلام [۱۲] و شکل نوشتاری کلمه هدف و کلمات مشابه در حوزه معنایی به عبارت ورودی را شناسایی نمودند. با ترکیب این چند شبکه، برای هر کلمه، یک امتیاز بر اساس میزان شباهت به عبارت ورودی در نظر گرفته شد. بدین ترتیب، شبیه‌ترین کلمات به عبارت ورودی تعیین گردید.

[۱۱] Bidirectional long short-term memory

[۱۲] Part of speech

فصل سوم

داده‌ها

این فصل به نحوه جمع‌آوری و اصلاح ساختار داده‌های موردنیاز برای انجام آزمایش‌ها می‌پردازد. در ابتدا، انواع داده‌های موردنیاز برای انجام آزمایش‌ها شرح داده می‌شوند. سپس نحوه جمع‌آوری داده‌ها از منابع زبانی مختلف و چالش‌های ناشی از انتخاب این داده‌ها بیان گردیده و شیوه تغییر ساختار آن‌ها ذکر می‌گردد. از داده‌های نهایی برای آموزش شبکه‌های عصبی و ارزیابی آن‌ها استفاده خواهد شد.

۳-۱ جمع‌آوری

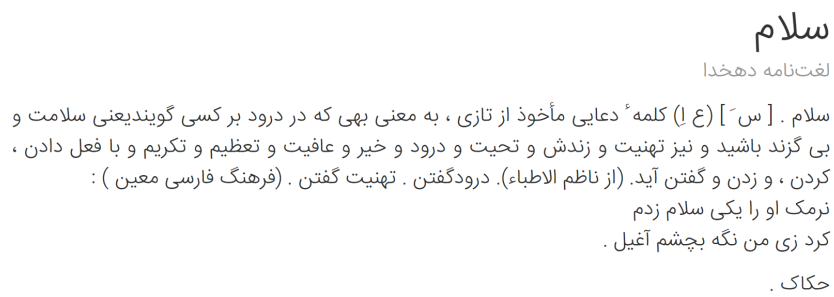
برای آموزش یک مدل که بتواند با دریافت یک عبارت توصیفی، کلمه‌ای را تحویل دهد، به داده‌هایی نیاز داریم که شامل کلمات و توصیف آن‌ها باشند. با توجه به محدودیت حافظه و زمان موردنیاز برای آموزش، تنها از بخشی از این داده‌ها استفاده خواهیم کرد. برای انتخاب این بخش از داده‌ها، از یک پیکره بزرگ استفاده می‌کنیم که از ترکیب چند پیکره دیگر پدید آمده است. علاوه بر آن، برای ارزیابی مدل‌ها به مجموعه کلمات مترادف هر کلمه نیاز خواهیم داشت. در این بخش، منابع مختلف استخراج داده‌ها معرفی می‌شود. برای هر منبع، نوع داده‌هایی که در آن وجود دارد و حجم آن‌ها ذکر می‌گردد.

۳-۱-۱ لغت‌نامه‌های فارسی

در این پژوهش از لغت‌نامه‌های عمید، معین و دهخدا استفاده شده است. مدخل‌های این لغت‌نامه‌ها از تارنمای واژه‌یاب [۱] توسط یک خزنده وب [۲] در قالب صفحات HTML دریافت شده‌اند. جدول ۳-۱ تعداد صفحات دریافت‌شده مربوط به هر لغت‌نامه را نشان می‌دهد. در هر مدخل، یک کلمه به همراه آواشناسی، الگوی تکیه، شمارگان هجا، برابر ابجاد، معانی آن و جزء کلام ذکر شده است. شکل ۳-۱ یک نمونه از مدخل‌های کلمات موجود در لغت‌نامه‌ها را نشان می‌دهد.

[۱] http://vajehyab.com

[۲] Web crawler


ویژگی‌های واژه

واژه: سلام
نقش دستوری: اسم
آواشناسی: salAm
الگوی تکیه: WS
شمارگان هجا: ۲
برابر ابجد: ۱۳۱

سلام

لغت‌نامه دهخدا

سلام . [س َ] (ع اِ) کلمهٔ دعایی مأخوذ از تازی ، به معنی بهی که در درود بر کسی گویندیعنی سلامت و بی گزند باشید و نیز تهنیت و زندش و تحیت و درود و خیر و عافیت و تعظیم و تکریم و با فعل دادن ، کردن ، و زدن و گفتن آید. (از ناظم الاطباء). درودگفتن . تهنیت گفتن . (فرهنگ فارسی معین) :
نرمک او را یکی سلام زدم
کرد زی من نگه بچشم آغیل .
حکاک .

شکل ۳-۱: بخشی از مدخل کلمه سلام در لغت‌نامه دهخدا، دریافت‌شده از تارنمای واژه‌یاب

لغت‌نامه	تعداد صفحات
عمید	۴۳۰۸۱
معین	۴۹۷۰۸۲
دهخدا	۱۲۹۸۲۶

جدول ۳-۱: تعداد صفحات دریافت‌شده از تارنمای واژه‌یاب به تفکیک لغت‌نامه

۳-۱-۲ ویکی‌پدیای فارسی

هر صفحه ویکی‌پدیای فارسی، شامل اطلاعاتی در مورد یک موجودیت [۳] فارسی است. مجموعه کلیه این صفحات، تحت عنوان پایگاه داده‌های ذخیره‌شده [۴] در ویکی‌مدیا [۵] ارائه می‌شود. در این پژوهش، از نسخه XML این صفحات (ارائه‌شده در تاریخ ۱۶ مارس سال ۲۰۲۰ میلادی) استفاده شده است. تعداد کل این صفحات ۶۰۹۸۵۰ عدد می‌باشد و هر صفحه XML حاوی ۴ تگ از جمله عنوان صفحه، عنوان بخش‌ها، متن بخش‌ها و پیوندهای درونی [۶] است.

۳-۱-۳ شبکه واژگانی فارسی

فارس نت [۷] یک شبکه واژگانی است که توسط شمس‌فرد و همکاران [۴۴] در آزمایشگاه پردازش زبان طبیعی دانشگاه شهیدبهشتی توسعه یافته است. نسخه سوم این شبکه واژگانی شامل اطلاعات مربوط به

[۳] Entity
[۴] Database dumps
[۵] https://dumps.wikimedia.org/backup-index.html
[۶] Interlinks
[۷] http://farsnet.nlp.sbu.ac.ir/

۱۰۰۰۰۰ کلمه فارسی می‌باشد. از طریق وب سرویس [۸] فارس‌نت [۹] می‌توان اطلاعات مربوط به کلمات را دریافت نمود.

۳-۱-۴ پیکره فارسی Uppsala

پیکره فارسی Uppsala [۴۳] که نسخه تغییریافته‌ای از پیکره بی‌جن‌خان [۴] می‌باشد، شامل ۲۷۰۴۰۲۸ واحد [۱۰] و برچسب جزء کلام برای هر کدام از آن‌ها است. پیکره اولیه که توسط بی‌جن‌خان و همکاران منتشر شده، از مجموعه اخبار روزانه جمع‌آوری شده است.

۳-۱-۵ پیکره همشهری

این پیکره شامل مجموعه اخبار روزنامه همشهری از سال ۱۳۷۵ الی ۱۳۸۱ است که از طریق تارنمای همین روزنامه دریافت شده است. در پیکره همشهری ۴۱۷۳۳۹ کلمه در قالب ۱۶۶۷۷۴ خبر مربوط به ۱۶ حوزه موجود است. پیکره همشهری در قالب یک فایل بزرگ شامل تمامی اخبار ارائه شده است. شکل ۳-۲ یک نمونه از اخبار موجود در این پیکره را نشان می‌دهد. به هر خبر یک شناسه اختصاص داده شده است. علاوه بر آن، تاریخ درج هر خبر و موضوع آن نیز ذکر شده است [۱].

.DID 4S2
.Date 75\04\30
.Cat elmfa
آیا بهتر نیست مدیران مدارس تعمیر و رنگ هرساله تخته سیاه را در سر فصل برنامه هاي کارتابستاني خود قرار دهند؟ در تابستان هر سال خستگي يك سال تحصيلي رااز روي تخته سياههاي مدارس بزداييم و با زدن رنگ و تعمير کردن آنها به استقبال سال تحصيلي جديدبرويم و بگذاريم اين وسيله آموزشي نيز در ابتداي سال، با ظاهري زيبا، پذيراي دانش آموزان در مدارس باشد. تخته سياه، وسيله اي موثر در ...

شکل ۳-۲: نمونه‌ای از یک خبر موجود در پیکره همشهری [۱]

[۸] Web service

[۹] http://farsnet.nlp.sbu.ac.ir/Site3/Modules/Public/WebAPI.jsp

[۱۰] Token

۳-۱-۶ فرهنگ جامع واژگان مترادف و متضاد زبان فارسی

در این پژوهش، برای ارزیابی عملکرد مدل‌ها، از مترادف کلمات استفاده شده است. بدین منظور، نسخه دیجیتال فرهنگ جامع واژگان مترادف و متضاد زبان فارسی [۵۴] ، در قالب یک فایل متنی از طریق تارنمای پیکره‌گان [۱۱] دریافت شد. نسخه اصلی این فرهنگ شامل ۱۵۰۰۰ مدخل در ۲۷۴۰۰ حوزه معنایی و ۱۳۵۰۰۰ واژه می‌باشد. در نسخه دیجیتال، در هر سطر، یک کلمه به همراه مجموعه کلمات مترادف آن درج شده است. شکل ۳-۳ چند سطر از این نسخه را نشان می‌دهد.

```
یزدان: آفریدگار، الله، ایزد، پروردگار، جان‌آفرین، خدا، رب & شیطان
یزدانی: الهی، ایزدی، ربانی، صمدانی & شیطانی
یزک: پیشقراول، دیده‌بان، دیده‌ور، قراول
یسار: ۱ چپ، میسره، یاسر ۲ استطاعت، تمول، توانگری ۳ شوم، نامیمون
```

شکل ۳-۳: چند نمونه از سطرهای نسخه دیجیتال فرهنگ جامع واژگان مترادف و متضاد زبان فارسی

[۱۱] https://www.peykaregan.ir/node/324?rid=252

۳-۲ چالش‌های ناشی از داده‌ها

استفاده از داده‌هایی که از منابع مختلف زبانی استخراج شده‌اند، چالش‌هایی را در مسیر طراحی فرهنگ واژگان وارونه به همراه می‌آورد. برخی از این چالش‌ها عبارتند از:

۳-۲-۱ تعارض میان معانی مختلف کلمات

گاهی اوقات یک کلمه معانی کاملا متفاوتی دارد. برای مثال کلمه «فرهنگ» را در نظر بگیرید. یک فارسی زبان می‌تواند این کلمه را به معنی «لغت‌نامه» استفاده کند (فرهنگ معین) یا آن را به معنی «آداب و رسوم یک قوم» به کار ببرد (فرهنگ ایل قشقایی). فرهنگ واژگان وارونه هوشمند باید تمامی معانی را برای هر کلمه بیاموزد.

۳-۲-۲ تغییرات زبان در طول زمان

با مقایسه معانی کلمات مشابه در لغت‌نامه‌های متفاوت، می‌توان به تغییرات زبانی که در طول سال‌های بسیار رخ داده است پی برد. به طور مثال، در مدخل کلمه «کرفس» در لغت‌نامه دهخدا عبارت زیر مشاهده می‌شود.

«رستنی باشد که از آن ترشی سازند یعنی در میان سرکه اندازند و خورند»

به عنوان بخشی از معنی این کلمه ذکر شده است. در این عبارت کلمه «رُستَن» به معنای «روییدن گیاه» به کار برده شده که مربوط به زبان پهلوی می‌باشد. نحوه به کار بردن افعال نیز تا حدودی نمایانگر نحوه نگارش قدیمی می‌باشد. برای همین کلمه در لغت‌نامه عمید - که بیش از ۳۰ سال پس از انتشار لغت‌نامه دهخدا منتشر شده است - عبارت زیر ذکر شده است.

«گیاهی علفی و دوساله، با ساقه‌های بلند و برگ‌های بریده که مصرف خوراکی دارد.»

این مثال نشان می‌دهد که با گذشت زمان، افراد علاوه بر کلمات متفاوت، از شیوه بیان متفاوتی نیز برای انتقال مفاهیم موردنظر خود استفاده می‌کنند. فرهنگ واژگان وارونه هوشمند باید قابلیت درک عبارات توصیفی قدیمی و جدید را دارا باشد تا بتواند به درستی نزدیک‌ترین کلمه به هر عبارت توصیفی را شناسایی کند.

۳-۳ آماده‌سازی

برای استفاده از داده‌های خام بدست‌آمده از لغت‌نامه‌ها که در قالب صفحات HTML تهیه شده‌اند، با پیش‌پردازش‌هایی آن‌ها را به مجموعه‌ای از داده‌های ساختاریافته تبدیل کردیم. همچنین، از هر صفحه ویکی‌پدیای فارسی، تعدادی عبارت توصیفی استخراج شد. علاوه بر آن، با استفاده از پیکره‌ها، یک رتبه‌بندی برای کلمات زبان فارسی بدست آمد. تعدادی عبارت توصیفی نیز از طریق شبکه واژگانی فارسی (فارس نت) استخراج شد. در نهایت، یک مجموعه از کلمات مترادف نیز ساخته شد. در این قسمت، جزئیات این مراحل شرح داده می‌شود.

۳-۳-۱ تغییر شکل داده‌های لغت‌نامه‌ها

برای استفاده از داده‌های موجود در صفحات HTML بدست‌آمده از تارنمای واژه‌یاب، عملیات زیر روی آن صفحات صورت گرفت.

حذف تگ‌های HTML

در این مرحله تنها بخشی از صفحه که مربوط به مدخل لغت‌نامه‌ها می‌باشد، استخراج گشت و کلیه قسمت‌های دیگر صفحات تارنمای واژه‌یاب حذف گردید.

حذف صفحات فاقد کاراکترهای فارسی یا عربی

برخی از صفحات تارنمای واژه‌یاب فاقد هرگونه محتوای فارسی یا عربی هستند. این موارد به طور کلی از داده‌ها حذف شد.

جداسازی معانی و اجزاء کلام

در این مرحله ابتدا معانی و اجزاء کلام از هر صفحه استخراج شد. مدخل هر کلمه شامل معانی مختلف آن است. برای هر معنی، جزء کلام مربوط به آن نیز درج شده است. برای هر صفحه (مدخل)، متن مربوط به هر معنی موجود در آن صفحه جدا گردید و به هر معنی یک شناسه نسبت داده شد. جزء کلام مربوط به هر معنی نیز استخراج گردید.

هر معنی را می‌توان با عبارات مختلفی توصیف نمود. به طور مثال برای کلمه «سلام» دو عبارات «درود گفتن» و «تهنیت گفتن» از یک معنی نمایندگی می‌کنند. این گونه عبارات نیز از یکدیگر تفکیک

شدند. پس از تغییر شکل داده‌های لغت‌نامه‌ها، اطلاعات مربوط به ۶۶۶۹۴۱ مدخل مانند نمونه ۳-۳-۱ به دست آمد.

نمونه ۳-۳-۱. نحوه جداسازی معانی و اجزاء کلام برای کلمه «دیوسالار»

متن بدست‌آمده از مراحل قبلی:

دیوسالار

(صفت) [قدیمی] ۱. دلیر؛ دلاور. ۲. دلیر و تندخو مانند دیو.

خروجی این مرحله:

کلمه: دیوسالار

جزء کلام کلیه معانی: صفت

معنی ۱:

جزء کلام مخصوص معنی ۱: ندارد

عبارت توصیفی ۱: دلیر

عبارت توصیفی۲: دلاور

معنی ۲:

جزء کلام مخصوص معنی ۲: ندارد

عبارت توصیفی ۱: دلیر و تندخو مانند دیو

۳-۳-۲ نرمال‌سازی [۱۲] کاراکترها

ابتدا لیستی از کاراکترهای موجود در متن اخبار همشهری، پیکره فارسی Uppsala، صفحات ویکی‌پدیای فارسی و همچنین کلمات و عبارات توصیفی حاصل از تغییر شکل داده‌های لغت‌نامه‌ها یا موجود در فرهنگ جامع واژگان مترادف و متضاد زبان فارسی بدست آمد. برخی از کاراکترها در صفحه کلید استاندارد فارسی موجود نبودند. برای این گونه کاراکترها، اگر معادل فارسی وجود داشت، آن معادل در هر کدام از منابع داده‌ای جایگزین کاراکتر اصلی شد و در غیر این صورت، کاراکتر حذف گردید. کاراکترهای انگلیسی و یا علائم خاص نیز کنار گذاشته شدند. علاوه بر آن، در برخی موارد عبارات توصیفی شامل علائم آوایی می‌شدند که به تلفظ بهتر کمک می‌کنند. از آنجایی که تعداد این علائم در کل داده‌ها بسیار کم بود،

[۱۲] Normalizing

عملا قابلیت استفاده از آن‌ها وجود نداشت. لذا آن‌ها را نیز حذف نمودیم. جدول ۳-۲ لیست کاراکترهای باقی‌مانده را نشان می‌دهد.

ء	ا	آ	ب	پ	ت	ث	
ج	چ	ح	خ	د	ذ	ر	
ز	ژ	س	ش	ص	ض	ط	
ظ	ع	غ	ف	ق	ک	گ	
ل	م	ن	و	ه	ی		

جدول ۳-۲: لیست کاراکترهای باقی‌مانده پس از نرمال‌سازی

۳-۳-۳ استخراج عبارات توصیفی از داده‌های ویکی‌پدیای فارسی

ویکی‌پدیای فارسی شامل هزاران مقاله است که هر کدام یک موجودیت را به زبان فارسی توصیف می‌کنند. اولین پاراگراف هر مقاله، معمولا بخش اصلی متن آن را تشکیل می‌دهد. از همین رو، از هر صفحه ویکی‌پدیا، پاراگراف اول آن را استخراج نمودیم و هر جمله از آن پاراگراف را به عنوان یک عبارت توصیفی در نظر گرفتیم.

نمونه ۳-۳-۲. نحوه استفاده از متن موجود در پاراگراف اول یک صفحه ویکی‌پدیا

متن صفحه:

خوشحالی

خوشحالی، خرسندی یا شادی (به انگلیسی: Happiness) یک حالت روانی است که در آن فرد احساس عشق، لذت، خوشبختی یا شاد بودن می‌کند. مسائل مختلفی همچون مسائل زیستی، روان‌شناختی یا دینی برای تعریف و دلیل خرسندی آورده شده‌است.

خروجی:

کلمه: خوشحالی

عبارت توصیفی ۱: خوشحالی، خرسندی یا شادی به انگلیسی یک حالت روانی است که در آن فرد احساس عشق، لذت، خوشبختی یا شاد بودن می‌کند

عبارت توصیفی ۲: مسائل مختلفی همچون مسائل زیستی، روان‌شناختی یا دینی برای تعریف و دلیل خرسندی آورده شده‌است.

پس از اجرای این مرحله، مجموعا اطلاعات مربوط به ۶۰۶۱۹۸ کلمه مانند نمونه ۳-۳-۲ بدست آمد. برای این کلمات، مجموعا ۲۷۱۴۴۰۷ عبارت توصیفی حاصل شد.

۳-۳-۴ استخراج متن اخبار از پیکره همشهری

در این پژوهش از میان اطلاعاتی که در پیکره همشهری موجود است، تنها از متن اخبار آن استفاده شده است. از همین رو، شناسه هر خبر، تاریخ درج آن و موضوع خبر حذف گردید و در نهایت یک فایل شامل متن کلیه اخبار بدست آمد.

۳-۳-۵ استخراج واحدها از پیکره فارسی Uppsala

برای استفاده از این پیکره، کلیه اجزاء کلام را از آن حذف نمودیم. در نهایت تعدادی واحد باقی ماند و در یک فایل متنی ذخیره شد.

۳-۳-۶ رتبه‌بندی کلمات

احتمال رخداد کلمات در یک پیکره بزرگ می‌تواند به عنوان تقریب خوبی از احتمال رخداد آن‌ها در یک زبان در نظر گرفته شود. بدین منظور ابتدا با الحاق داده‌های بدست‌آمده از منابع، یک پیکره بزرگ شامل موارد زیر تشکیل شد.

- عبارات توصیفی موجود در لغت‌نامه‌ها
- متن کامل صفحات ویکی‌پدیای فارسی
- متن اخبار موجود در پیکره همشهری
- واحدهای موجود در پیکره فارسی Uppsala

برای هر کلمه موجود در لغت‌نامه‌ها، تعداد رخداد آن کلمه در این پیکره محاسبه گردید. بر این اساس، به هر کلمه یک رتبه نسبت داده شد؛ به طوری که کلمات با تعداد رخداد بیشتر، در رتبه‌های بالاتر قرار گرفتند. این رتبه‌بندی حاوی ۱۳۸۵۰۳۰ واحد مجزا می‌باشد.

۳-۳-۷ استخراج عبارات توصیفی از فارس نت

با استفاده از وب سرویس فارس نت، می‌توان به عبارات توصیفی برای هر کلمه موجود در آن، دست یافت. در این پژوهش، برای ۳۰۰۰۰ کلمه نخست از نظر رتبه، عبارات توصیفی جستجو شد و اطلاعات مربوط به ۲۱۱۹۴ کلمه استخراج گشت. کاراکترهای موجود در این عبارات نیز نرمال شدند.

۳-۳-۸ ترکیب مجموعه کلمات مترادف

گاهی اوقات یک کلمه به عنوان معنی کلمه‌ای دیگر در لغت‌نامه‌ها درج شده است. در این گونه موارد، دو کلمه را مترادف فرض می‌کنیم. پس از ترکیب این داده‌ها با اطلاعات موجود در فرهنگ جامع واژگان مترادف و متضاد زبان فارسی [۵۴] مجموعه کامل‌تری از مترادف‌ها بدست می‌آید. بدین ترتیب، کلمات مترادف مشخص می‌شوند. پس از این مرحله، برای ۲۵۱۲۷۳ کلمه، کلمات مترادف ذخیره شد.

۳-۳-۹ ترکیب کلمات و عبارات توصیفی منابع مختلف

در بخش ۳-۳-۱ معانی و اجزاء کلام هر مدخل موجود در لغت‌نامه‌ها را جدا نمودیم. علاوه بر آن، به هر معنی کلمه، یک شناسه نسبت دادیم. بدین ترتیب برای معنی n-ام یک کلمه مانند w، عباراتی توصیفی مانند $d_1, d_2, \ldots, d_t$ از لغت‌نامه s استخراج شد. در این مرحله داده‌های مربوط به کلمه w را به شکل چهارتایی‌های (d_i, w, n, s) ذخیره می‌کنیم.

همچنین، اگر $d'_1, d'_2, \ldots, d'_m$ عبارات در وصف کلمه w' باشند که مطابق بخش ۳-۳-۳ از ویکی‌پدیای فارسی استخراج شده‌اند، برای همگونی با داده‌های لغت‌نامه‌ها به آن‌ها شناسه ۰ نسبت می‌دهیم و هر یک از آن‌ها را به شکل $(d'_j, w', 0, \text{wiki})$ ذخیره می‌کنیم.

به طور مشابه اگر $d''_1, d''_2, \ldots, d''_p$ عبارات توصیفی مربوط به کلمه w'' باشند که طبق بخش ۳-۳-۷ از فارس‌نت استخراج شده‌اند، هر یک از آن‌ها را به شکل $(d''_k, w'', 0, \text{fnet})$ ذخیره می‌کنیم.

با ترکیب این چهارتایی‌های مرتب، مجموعا ۳۸۳۳۴۵۶ چهارتایی مرتب حاصل می‌شود. در ادامه این پایان‌نامه‌ها این چهارتایی‌ها را «اطلاعات تفکیک‌شده کلمات» می‌نامیم.

۳-۴ جمع‌بندی

در این فصل، منابع مختلف استخراج داده‌ها به همراه ساختار اولیه آن‌ها ذکر شد. چالش‌های استفاده از آن‌ها برای آموزش یک مدل فرهنگ واژگان وارونه مطرح گردید و نحوه تغییر ساختار آن‌ها به فرم قابل استفاده برای این پژوهش نیز بیان شد. در نهایت یک روش رتبه‌بندی برای کلمات بیان شد و نحوه دست‌یابی به مجموعه کلمات مترادف یک کلمه، ذکر گردید. پس از کلیه این مراحل، سه مجموعه داده بدست آمد که عبارتند از:

- اطلاعات تفکیک‌شده کلمات
  - تعداد نمونه‌ها: ۳۸۳۳۴۵۶
  - هر نمونه شامل: یک کلمه، یک عبارت توصیفی، یک شناسه و یک منبع
  - نحوه تشکیل نمونه‌ها: ترکیب داده‌های لغت‌نامه‌ها، ویکی‌پدیای فارسی و فارس‌نت
  - نوع داده: چهارتایی مرتب
- مجموعه کلمات مترادف
  - تعداد کلمات حاوی مجموعه کلمات مترادف: ۲۵۱۲۷۳
  - نحوه تشکیل نمونه‌ها: ترکیب داده‌های لغت‌نامه‌ها و فرهنگ جامع واژگان مترادف و متضاد فارسی
  - نوع داده: یک شیء از نوع dictionary در زبان پایتون که هر کلمه را به یک مجموعه (شیء از نوع set) نگاشت می‌کند.
- رتبه‌بندی کلمات
  - تعداد واحدهای رتبه‌بندی‌شده: ۱۳۸۵۰۳۰
  - نوع داده: یک لیست مرتب

فصل چهارم

معماری

این فصل به بررسی مدل‌های پیشنهادی برای حل مسئله می‌پردازد. برای درک نحوه عملکرد این مدل‌ها، آشنایی با برخی از مفاهیم مرتبط با شبکه‌های عصبی ضروری است. از همین رو، ابتدا این مفاهیم شرح داده می‌شود؛ سپس معماری مدل‌های پیشنهادی تشریح می‌گردد.

۴-۱ پیش‌زمینه

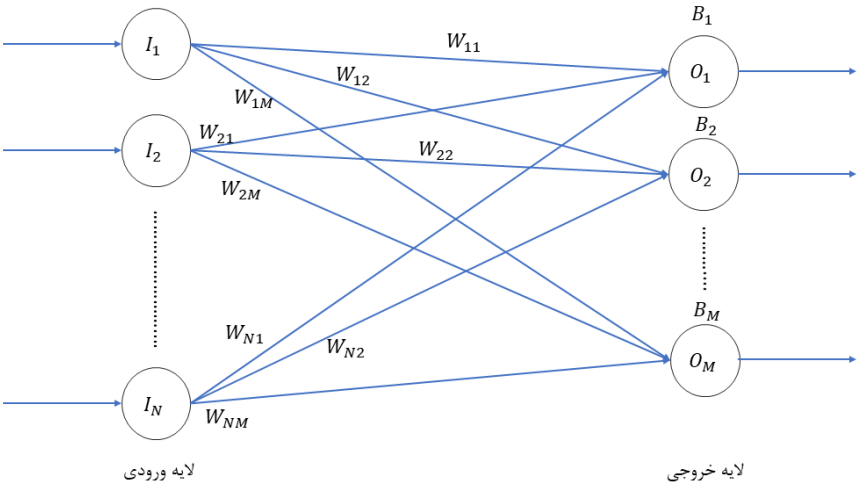
در این قسمت، مفاهیم اولیه مورد نیاز برای درک مدل‌های پیشنهادی بیان می‌گردد.

۴-۱-۱ شبکه تماما متصل [۱]

یک شبکه تماما متصل شامل یک لایه ورودی و یک لایه خروجی می‌باشد. اگر مقادیر نورون‌های لایه ورودی و خروجی را به ترتیب با نمادهای $I_1, I_2, \ldots, I_N$ و $O_1, O_2, \ldots, O_M$ نمایش دهیم، داریم:

$$\forall j \in \{1, 2, \ldots, M\} \quad O_j = \phi(\sum_{k=1}^{N} I_k W_{kj} + B_j) \tag{۴-۱}$$

در عبارت فوق، W ماتریسی شامل وزن‌های میان نورون‌های لایه ورودی و خروجی، B بردار بایاس [۲] و ϕ تابع فعال‌سازی [۳] می‌باشد. شکل ۴-۱ معماری این شبکه را در قالب یک گراف نمایش می‌دهد.

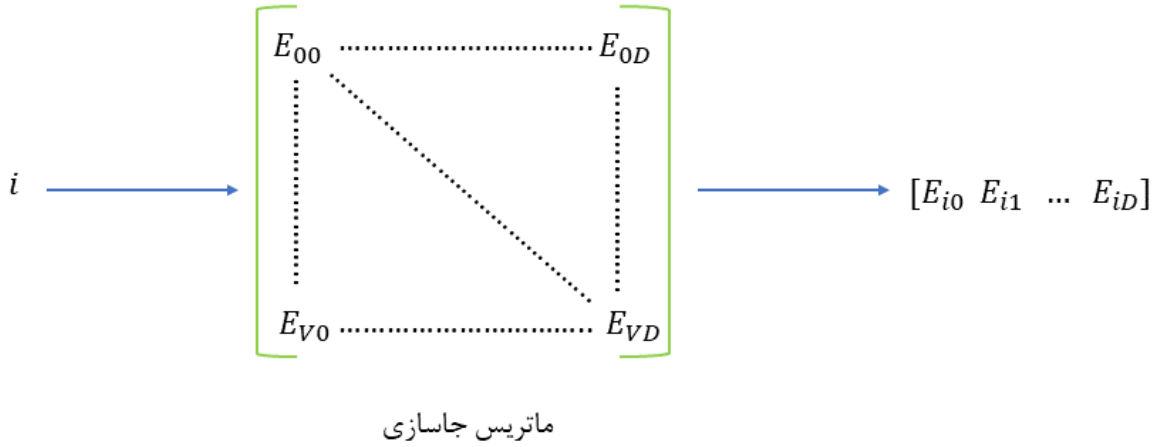


شکل ۴-۱: شبکه تماما متصل

[۱] Fully-connected network
[۲] Bias vector
[۳] Activation Function

۴-۱-۲ لایه جاسازی [۴]

لایه جاسازی شامل یک ماتریس جاسازی است. این لایه یک عدد صحیح نامنفی مانند i را دریافت می‌کند و سطر i-ام ماتریس مذکور را تحویل می‌دهد. شکل ۴-۲ نحوه عملکرد این لایه را نشان می‌دهد.



شکل ۴-۲: لایه جاسازی

۴-۱-۳ شبکه عصبی بازگشتی [۵]

این نوع شبکه‌ها، از مدل‌های یادگیری با نظارت هستند که به دلیل دارابودن ارتباطات بازگشتی، قادر به یادگیری ویژگی‌های زمانی یک مجموعه داده می‌باشند [۱۴]. تا کنون از آن‌ها در حوزه‌های مختلفی از جمله یادگیری مدل‌های زبانی [۶] [۳۱] ، پیش‌بینی بازدهی سهام [۳۹] و پیش‌بینی امواج اقیانوس [۲۶] استفاده شده است.

یک شبکه عصبی بازگشتی ساده شامل سه لایه است که به ترتیب لایه ورودی، لایه مخفی بازگشتی و لایه خروجی نامیده و با نمادهای V ، H و O نشان داده می‌شوند. اگر این لایه‌ها به ترتیب دارای N ، M و P نورون باشند، وضعیت آن‌ها در زمان t را به ترتیب با نمادهای $V_t = (x_1, x_2, \ldots, x_N)$ ، $H_t = (h_1, h_2, \ldots, h_M)$ و $O_t = (o_1, o_2, \ldots, o_P)$ نشان دهید.

لایه ورودی دنباله‌ای از بردارها مانند $(V_1, V_2, \ldots, V_T)$ را در طول زمان دریافت می‌کند؛ در یک شبکه عصبی بازگشتی تماما متصل، نورون‌های لایه ورودی به نورون‌های لایه مخفی متصل هستند. این اتصال‌ها

[۴] Embedding
[۵] Recurrent neural network
[۶] Language models

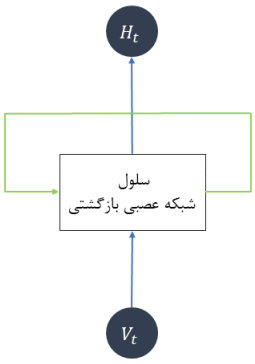
توسط یک ماتریس وزن [۷] مانند W_{VH} مدل می‌شوند. نورون‌های لایه مخفی نیز به یکدیگر در طول زمان متصل هستند و این اتصال با ماتریس دیگری مانند W_{HH} تعریف می‌گردد. در نظر بگیرید $f_H(.)$ تابع فعال‌سازی لایه مخفی و B_H بردار بایاس نورون‌های آن باشد. مقادیر این نورون‌ها در زمان t به شکل زیر محاسبه می‌گردد.

$$H_t = f_H(W_{VH}V_t + W_{HH}H_{t-1} + B_H) \quad (۲-۴)$$

ارتباط میان نورون‌های لایه مخفی و لایه خروجی را نیز با ماتریس W_{HO} ، تابع فعال‌سازی لایه خروجی را با $f_O(.)$ و بردار بایاس نورون‌های لایه خروجی را با B_O نشان دهید. وضعیت لایه خروجی در زمان t از طریق رابطه زیر بدست می‌آید:

$$O_t = f_O(W_{HO}H_t + B_O) \quad (۳-۴)$$

از آنجایی که ورودی به صورت متوالی در گذر زمان به شبکه داده می‌شود، مراحل محاسبات نیز در گذر زمان $t \in \{1, 2, \ldots, T\}$ تکرار خواهند شد [۴۱]. شکل ۴-۳ معماری این شبکه را در یک لحظه از زمان مانند t نشان می‌دهد.



شکل ۴-۳: شبکه عصبی بازگشتی ساده

[۷] Weight matrix

۴-۱-۴ حافظه کوتاه-مدت ماندگار [۸]

با وجود توانایی یادگیری ویژگی‌های زمانی، شبکه‌های عصبی بازگشتی در یادگیری وابستگی‌های زمانی بلندمدت با پدیده‌های محو گرادیان [۹] و یا انفجار آن [۱۰] روبرو می‌شوند. برای حل این مشکل، Hochreiter و Schmidhuber [۲۲] شبکه عصبی بازگشتی جدیدی ابداع کردند و آن را حافظه کوتاه-مدت ماندگار نامیدند. مانند بخش قبلی، فرض کنید ورودی به صورت بردارهایی مانند V_t در طول زمان $t \in \{۱, ۲, \dots, T\}$ به شبکه ارائه می‌شود. این شبکه دارای ۶ لایه مخفی مانند P، I، F، O، M و H است که مقادیر آن‌ها در زمان t از طریق روابط زیر بدست می‌آید.

$$P_t = W_{VP}V_t + W_{HP}H_{t-۱} + B_P \tag{۴-۴}$$

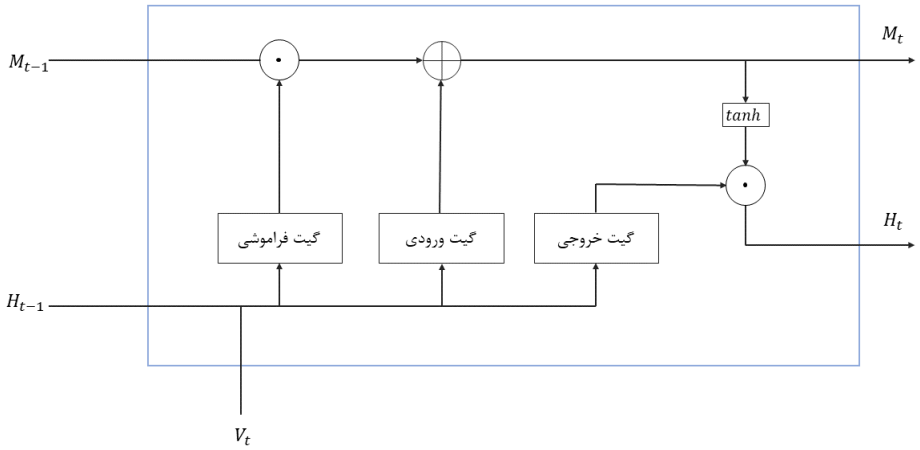
$$\forall L \in \{I, F, O\} \quad L_t = \frac{۱}{۱ + e^{-(W_{VL}V_t + W_{HL}H_{t-۱} + B_L)}} \tag{۴-۵}$$

$$M_t = P_t \odot I_t + M_{t-۱} \odot F_t \tag{۴-۶}$$

$$H_t = O_t \odot tanh(M_t) \tag{۴-۷}$$

در روابط بالا، $\odot$ نماد ضرب مولفه به مولفه است. گاهی به لایه‌های I، F و O به ترتیب گیت ورودی [۱۱]، گیت فراموشی [۱۲] و گیت خروجی [۱۳] گفته می‌شود. خروجی لایه H در هر لحظه از زمان مانند t، خلاصه‌ای از اطلاعات واردشده به حافظه کوتاه-مدت ماندگار تا لحظه t را در بر دارد [۲۱]. گاهی این لایه را حالت مخفی [۱۴] می‌نامند. شکل ۴-۴ معماری شبکه را در یک لحظه از زمان مانند t نشان می‌دهد.

[۸] Long short-term memory
[۹] Vanishing gradient
[۱۰] Exploding gradient
[۱۱] Input gate
[۱۲] Forget gate
[۱۳] Output gate
[۱۴] Hidden state



شکل ۴-۴: معماری یک شبکه حافظه کوتاه-مدت ماندگار در زمان t

۴-۱-۵ حافظه کوتاه-مدت ماندگار دوسویه [۱۵]

این نوع شبکه‌ها، متشکل از دو زیرشبکه از نوع حافظه کوتاه-مدت ماندگار هستند که یکی ورودی‌ها را از ابتدا به انتها و دیگری از انتها به ابتدا در طول زمان دریافت می‌کند. خروجی این شبکه در هر زمان مانند t از ترکیب خروجی‌های دو زیرشبکه در آن زمان بدست می‌آید. این ترکیب می‌تواند به شکل جمع دو مقدار یا الحاق آن‌ها باشد [۴۲].

۴-۱-۶ لایه توجه [۱۶]

لایه توجه، یک تابع توجه را پیاده‌سازی می‌کند. تابع توجه، نگاشتی از یک پرسش [۱۷] و یک مجموعه از دوتایی‌های (کلید و مقدار) به یک خروجی است که در آن پرسش، کلیدها، مقادیر و خروجی همگی بردار هستند. این بردارها در کاربردهای مختلف، معانی مختلفی دارند؛ به طور مثال، در پردازش زبان طبیعی، ممکن است هر کدام نمایش برداری یک کلمه باشند. خروجی لایه توجه، یک مجموع وزن‌دار از مقادیر است؛ به طوری که هر وزن اختصاص‌یافته به یک مقدار، توسط یک تابع امتیازدهی از پرسش و کلید متناظر آن پرسش، محاسبه می‌شود.

فرض کنید کلیه بردارهای پرسش، در یک ماتریس پرسش مانند $Q \in \mathbb{R}^{m \times d_k}$ قرار گرفته باشد. به طور مشابه، فرض کنید $K \in \mathbb{R}^{n \times d_k}$ و $V \in \mathbb{R}^{n \times d_v}$ ماتریس‌هایی شامل بردارهای کلید و مقدار هستند.

[۱۵] Bidirectional long short-term memory
[۱۶] Attention layer
[۱۷] Query

در حالت کلی می‌توان تابع توجه را بر اساس Q، K و V به شکل زیر تعریف کرد [۵۰].

$$Attention(Q, K, V) = softmax(f(Q, K))V \quad (۸-۴)$$

در رابطه فوق، f یک تابع امتیازدهی است. همچنین تابع $softmax$ برای هر بردار مانند $X = (x_۱, x_۲, \ldots, x_n)$ به شکل زیر تعریف می‌گردد:

$$softmax(X) = (\frac{e^{x_۱}}{\sum_{i=۱}^{n} e^{x_i}}, \frac{e^{x_۲}}{\sum_{i=۱}^{n} e^{x_i}}, \ldots, \frac{e^{x_n}}{\sum_{i=۱}^{n} e^{x_i}}) \quad (۹-۴)$$

در هنگام آموزش شبکه‌های عصبی، هر دسته از داده‌ها به صورت یک آرایه چندبعدی به شبکه وارد می‌شوند. این آرایه‌های چندبعدی را تنسور می‌نامیم. اگر ورودی تابع $softmax$ یک تنسور باشد، این تابع روی آخرین بُعد آن تنسور عمل خواهد کرد.

توجه از نوع ضرب نقطه‌ای مقیاس‌پذیر [۱۸]

در توجه از نوع ضرب نقطه‌ای مقیاس‌پذیر، مقدار تابع توجه به شکل زیر تعریف می‌گردد.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (۱۰-۴)$$

در رابطه (۴-۱۰) نماد T نشان‌دهنده عمل ترانهاده است و d_k بُعد هر یک از سطرهای ماتریس بردارهای کلید را نشان می‌دهد.

توجه جمع‌شونده [۱۹]

در این نوع توجه که توسط Bahdanau و همکاران [۲] پیشنهاد گردید ، مقدار توجه از طریق رابطه زیر محاسبه می‌گردد.

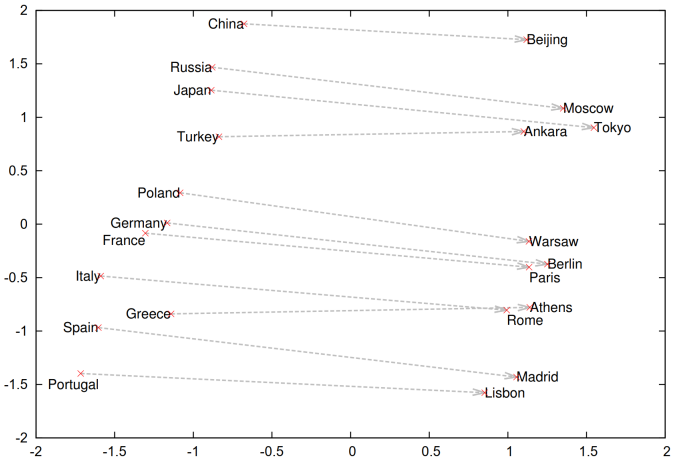
$$Attention(Q, K, V) = softmax(tanh(Q + K))V \quad (۱۱-۴)$$

[۱۸] Scaled dot-product attention

[۱۹] Additive attention

۴-۱-۷ بازنمایی برداری کلمات

در سال ۲۰۰۳ Bengio و همکاران [۳] با آموزش یک شبکه عصبی برای تخمین احتمال رخداد دنباله‌ای از کلمات، به نمایشی از کلمات در فضای بردارهای حقیقی-مقدار دست یافتند. آن‌ها به هر کلمه یک بردار در فضای $\mathbb{R}^m$ نسبت دادند. Mikolov و همکاران در سال ۲۰۱۳ مدل‌های دیگری را تحت نام Word2Vec برای بازنمایی برداری کلمات ارائه نمودند که قادر به یادگیری ویژگی‌های معنایی [۲۰] و نحوی زبان نیز می‌باشند [۲۹]. برای تجسم نتایج، آن‌ها بردارهای بدست‌آمده را به فضای دوبعدی تصویر کردند [۳۰]. شکل ۴-۵ که تصویر برخی از این بردارها در فضای دوبعدی را نشان می‌دهد، حاکی از آن است که بردار کلمات مربوط به کشورها و پایتخت آن‌ها، فواصل مشابهی دارند.



شکل ۴-۵: تصویر نمایش کلمات در فضای ۲-بعدی در مدل Mikolov و همکاران [۲۹]

مدل Mikolov و همکاران برای هر کلمه موجود در داده‌های آموزشی، یک بردار ارائه می‌کند؛ اما قادر به ارائه بردار برای کلمات جدید نیست. در سال ۲۰۱۷ Bojanowski و همکاران [۶] با در نظر گرفتن هر کلمه به عنوان سبدی از n-تایی های کاراکتری [۲۱] و ارائه بردار برای هر کدام از این n-تایی‌ها، مدلی را ارائه نمودند که ضعف مذکور را در موارد متعددی پوشش می‌دهد. آن‌ها به دلیل سرعت بالای آموزش مدل خود و توانایی آن در دسته‌بندی متون، آن را FastText نامیدند.

در سال ۲۰۱۸، Grave و همکاران [۱۸] با استفاده از همان معماری، مدل‌هایی را برای بازنمایی کلمات ۱۵۷ زبان از جمله فارسی منتشر کردند. در ادامه این پایان‌نامه، مدل ارائه‌شده آن‌ها برای زبان فارسی را با نماد FT نشان می‌دهیم. هادی‌فر و ممتازی [۱۹] در همان سال با بهره‌گیری از هر دو معماری

[۲۰] Semantic feature

[۲۱] Bag of character n-grams

Word2Vec و FastText مدل‌هایی را برای بازنمایی برداری کلمات فارسی ارائه نمودند. از میان آن مدل‌ها، دو مورد با استفاده از معماری Word2Vec و پیکره‌های همشهری و ویکی‌پدیای فارسی آموزش داده شده‌اند. در این پژوهش، به این مدل‌ها به ترتیب با نام‌های HAM-W2V و WIKI-W2V اشاره می‌کنیم.

۴-۱-۸ بیش‌برازش [۲۲]

اگر عملکرد یک مدل روی داده‌های آموزشی بسیار خوب و روی داده‌های آزمایشی به مراتب ضعیف‌تر باشد، اصطلاحا گوییم پدیده بیش‌برازش رخ داده است. در این حالت، مدل ویژگی‌هایی را آموخته است که مخصوص داده‌های آموزشی هستند و قابلیت تعمیم‌پذیری [۲۳] مناسبی ندارد. برای جلوگیری از وقوع پدیده بیش‌برازش، روش‌های زیادی وجود دارد که در ادامه به بررسی سه مورد از آن‌ها می‌پردازیم.

توقف زودهنگام [۲۴]

در این روش، از دو مجموعه داده آموزشی و اعتبارسنجی استفاده می‌شود. ابتدا با استفاده از داده‌های آموزشی، یک مدل آموزش داده می‌شود. پس از هر مرحله [۲۵] آموزش مدل، خطا روی داده‌های اعتبارسنجی با مرحله قبل مقایسه می‌شود. به محض افزایش خطا نسبت به مرحله قبلی، آموزش متوقف می‌گردد و وزن‌های مرحله قبل جایگزین وزن‌های کنونی می‌شوند. در این روش، فرض می‌شود خطا روی داده‌های اعتبارسنجی، تقریب مناسبی از خطا روی داده‌های آزمایشی می‌باشد [۳۸].

تنزل وزن [۲۶]

در این روش، تمامی متغیرهای قابل یادگیری مدل، در یک بردار وزن مانند $W = (w_1, w_2, \ldots, w_k)$ قرار می‌گیرند. اگر تابع هزینه [۲۷] مدل بدون استفاده از روش تنزل وزن با نماد $E_0(W)$ نشان داده شود،

[۲۲] Overfit
[۲۳] Generalization
[۲۴] Early stopping
[۲۵] Epoch
[۲۶] Weight decay
[۲۷] Cost function

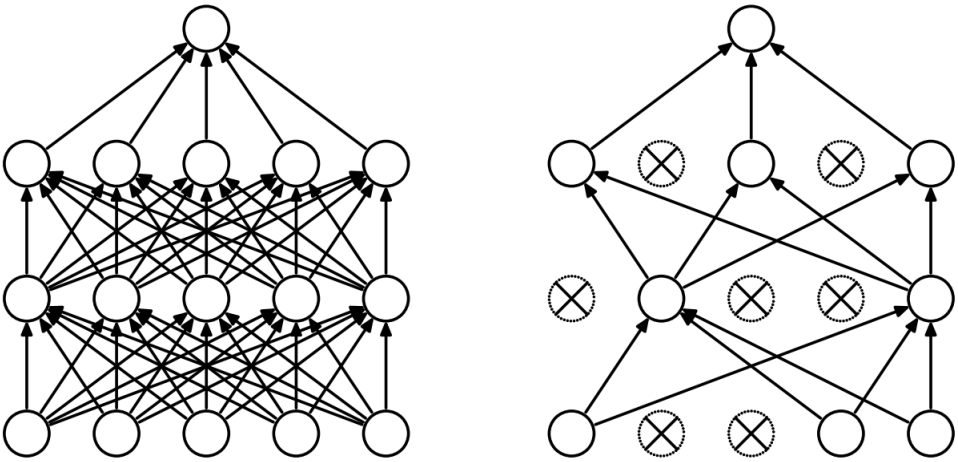
تابع هزینه جدید ($E(W)$) با در نظر گرفتن این روش به شکل زیر خواهد بود.

$$E(W) = E_{\circ}(W) + \frac{۱}{۲}\lambda \sum_{i=۱}^{k} w_i^۲ \qquad (۱۲\text{-}۴)$$

در رابطه فوق، λ یک پارامتر دلخواه است که میزان جریمه وزن‌های بزرگ را معین می‌کند [۲۴].

حذف تصادفی [۲۸]

این روش بر مبنای حذف موقتی نورون‌ها از شبکه عمل می‌کند. منظور از حذف یک نورون، حذف اتصالات ورودی به آن یا خروجی از آن است. انتخاب نورون‌ها برای حذف، تصادفی صورت می‌گیرد. بدین منظور، به هر نورون یک احتمال حذف مانند p مستقل از بقیه نورون‌ها اختصاص می‌یابد [۴۶]. شکل ۴-۶ نحوه تغییر اتصالات یک شبکه پس از به کارگیری روش حذف تصادفی را نشان می‌دهد.



شکل ۴-۶: سمت چپ، یک شبکه عصبی و سمت راست، همان شبکه پس از حذف تصادفی [۴۶]

[۲۸] Dropout

۴-۲ معماری مدل‌های پیشنهادی

این بخش به بررسی معماری مدل‌های پیشنهادی برای حل مسئله اختصاص یافته است. در کلیه مدل‌ها، دنباله‌ای از اعداد به عنوان ورودی پذیرفته می‌شود. سپس این دنباله وارد لایه جاسازی شده و به مجموعه‌ای از بردارها تبدیل می‌گردد. تفاوت معماری مدل‌های پیشنهادی با یکدیگر، در لایه‌های بعدی می‌باشد که در ادامه شرح داده می‌شود.

در اولین تلاش برای حل مسئله، مدلی ساده بر پایه میانگین‌گیری از بردارها ارائه خواهیم نمود. در این مدل ترتیب رخداد بردارها در نظر گرفته نمی‌شود. از آنجایی که عملیات میانگین‌گیری هزینه محاسباتی بسیار کمی دارد، در صورتی که مدل به نتایج قابل قبولی دست پیدا کند، نیازی به استفاده از مدل‌های پیچیده‌تر نخواهیم داشت.

سپس یک حافظه کوتاه-مدت ماندگار را جایگزین عملیات میانگین‌گیری خواهیم کرد. این نوع شبکه عصبی قادر به مدل‌کردن ترتیب رخداد بردارها می‌باشد؛ به همین دلیل، از لحاظ تئوری انتظار می‌رود حالت مخفی آن، اطلاعات ورودی در طول زمان را به خوبی خلاصه‌سازی کند.

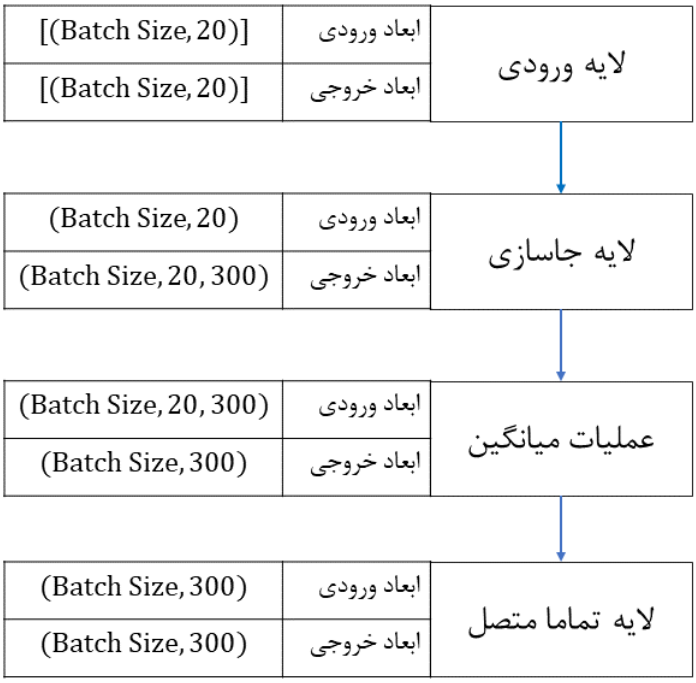
معمولا اجزاء مختلف اطلاعات ورودی، نقش یکسانی در تولید خروجی ندارند. برای لحاظ‌کردن اهمیت متفاوت این اجزاء، یک لایه توجه را به مدل اضافه خواهیم کرد. در نهایت برای جلوگیری از وقوع پدیده محو گرادیان، قابلیت پردازش دوسویه را به حافظه کوتاه-مدت ماندگار می‌افزاییم؛ با این امید که اجزاء اطلاعاتی متناظر آخرین لحظات زمان، نادیده گرفته نشوند.

۴-۲-۱ مدل سبد کلمات

در این مدل، ابتدا میانگین بردارهای دریافتی از لایه جاسازی محاسبه می‌گردد. سپس میانگین بدست‌آمده، تحت تاثیر یک لایه تماما متصل به یک بردار دیگر از همان بُعد تصویر می‌گردد. این بردار به عنوان خروجی نهایی مدل در نظر گرفته می‌شود. پارامترهای این مدل در جدول ۴-۱ ذکر شده است و شکل ۴-۷ معماری آن را نشان می‌دهد.

اندازه دنباله ورودی	تعداد نورون‌های لایه تماما متصل
۲۰	۳۰۰
بعد بردارها در ماتریس جاسازی	تابع فعال‌سازی لایه تماما متصل
۳۰۰	تانژانت هایپربولیک

جدول ۴-۱: پارامترهای مدل سبد کلمات



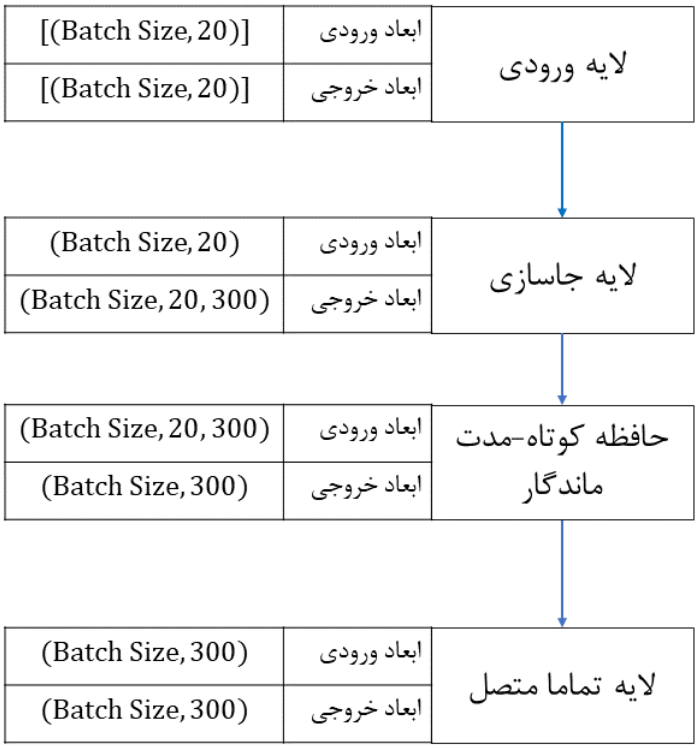
شکل ۴-۷: معماری مدل سبد کلمات

۴-۲-۲ مدل حافظه کوتاه-مدت ماندگار

در این مدل، بردارهای دریافتی از لایه جاسازی وارد حافظه کوتاه-مدت ماندگار می‌شوند. سپس آخرین حالت مخفی این لایه (H_T) وارد یک لایه تماما متصل می‌شود و این لایه، آن را به یک بردار از همان بُعد تصویر می‌کند. خروجی مدل، همان خروجی لایه تماما متصل است. جدول ۴-۲ پارامترهای مدل و شکل ۴-۸ معماری آن را نشان می‌دهد.

اندازه دنباله ورودی	تعداد نورون‌های لایه تماما متصل
۲۰	۳۰۰
بعد بردارها در ماتریس جاسازی	تابع فعال‌سازی لایه تماما متصل
۳۰۰	تانژانت هایپربولیک
تعداد نورون‌های حافظه کوتاه-مدت ماندگار	
۳۰۰	

جدول ۴-۲: پارامترهای مدل حافظه کوتاه-مدت ماندگار



شکل ۴-۸: معماری مدل حافظه کوتاه-مدت ماندگار

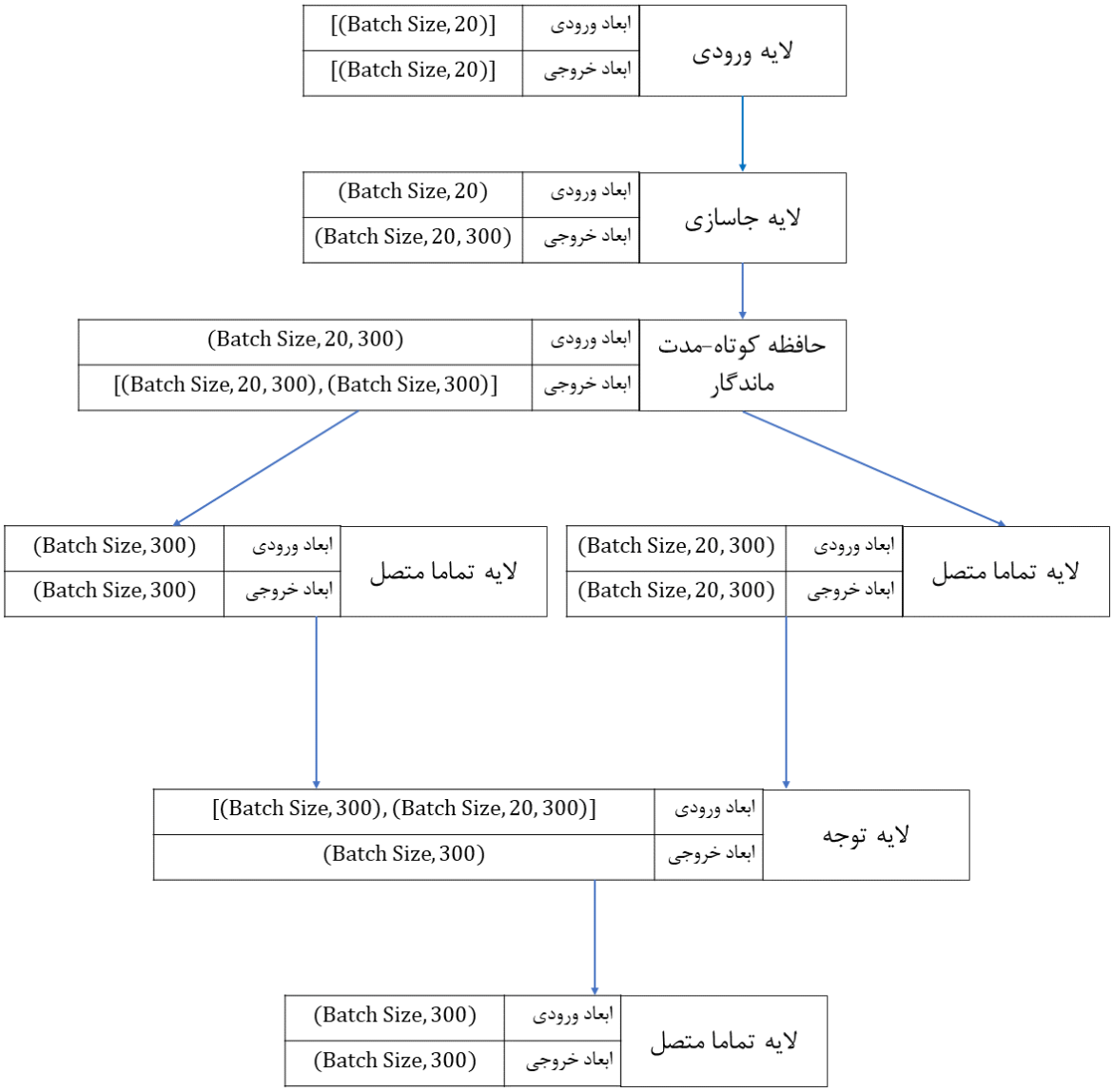
۴-۲-۳ مدل حافظه کوتاه-مدت ماندگار با توجه

در این مدل، ابتدا بردارها وارد یک حافظه کوتاه-مدت ماندگار می‌شوند. حافظه کوتاه-مدت ماندگار، حالت مخفی در هر لحظه از زمان را محاسبه می‌کند. آخرین حالت مخفی وارد یک لایه تماما متصل می‌شود. این لایه تماما متصل، یک تصویر از آخرین حالت مخفی را محاسبه می‌کند. این تصویر به عنوان پرسش برای لایه توجه به کار گرفته می‌شود.

از طرفی، به طور جداگانه، کلیه حالت‌های مخفی نیز وارد یک لایه تماما متصل می‌شوند. لذا تصاویری از حالت‌های مخفی در طول زمان نیز بدست می‌آید. این تصاویر به عنوان کلیدها و مقادیر وارد لایه توجه می‌شوند. لایه توجه با استفاده از پرسش (تصویر آخرین حالت مخفی) و مجموعه کلیدها و مقادیر (تصاویر مربوط به حالت مخفی در گذر زمان) میزان توجه به حالت مخفی در هر لحظه از زمان را محاسبه می‌کند. سپس یک مجموع وزن‌دار از حالت‌های مخفی در گذر زمان (یک بردار) را به عنوان خروجی تحویل می‌دهد. این بردار وارد یک لایه تماما متصل می‌گردد و به برداری از همان بُعد تصویر می‌شود. بردار نهایی، خروجی مدل است. جدول ۴-۳ پارامترهای مدل و شکل ۴-۹ معماری آن را نشان می‌دهد.

اندازه دنباله ورودی	تعداد نورون‌های هر لایه تماما متصل
۲۰	۳۰۰
بعد بردارها در ماتریس جاسازی	تابع فعال‌سازی هر لایه تماما متصل
۳۰۰	تانژانت هایپربولیک
تعداد نورون‌های حافظه کوتاه-مدت ماندگار	
۳۰۰	

جدول ۴-۳: پارامترهای مدل حافظه کوتاه-مدت ماندگار با توجه



شکل ۴-۹: معماری مدل حافظه کوتاه-مدت ماندگار با توجه

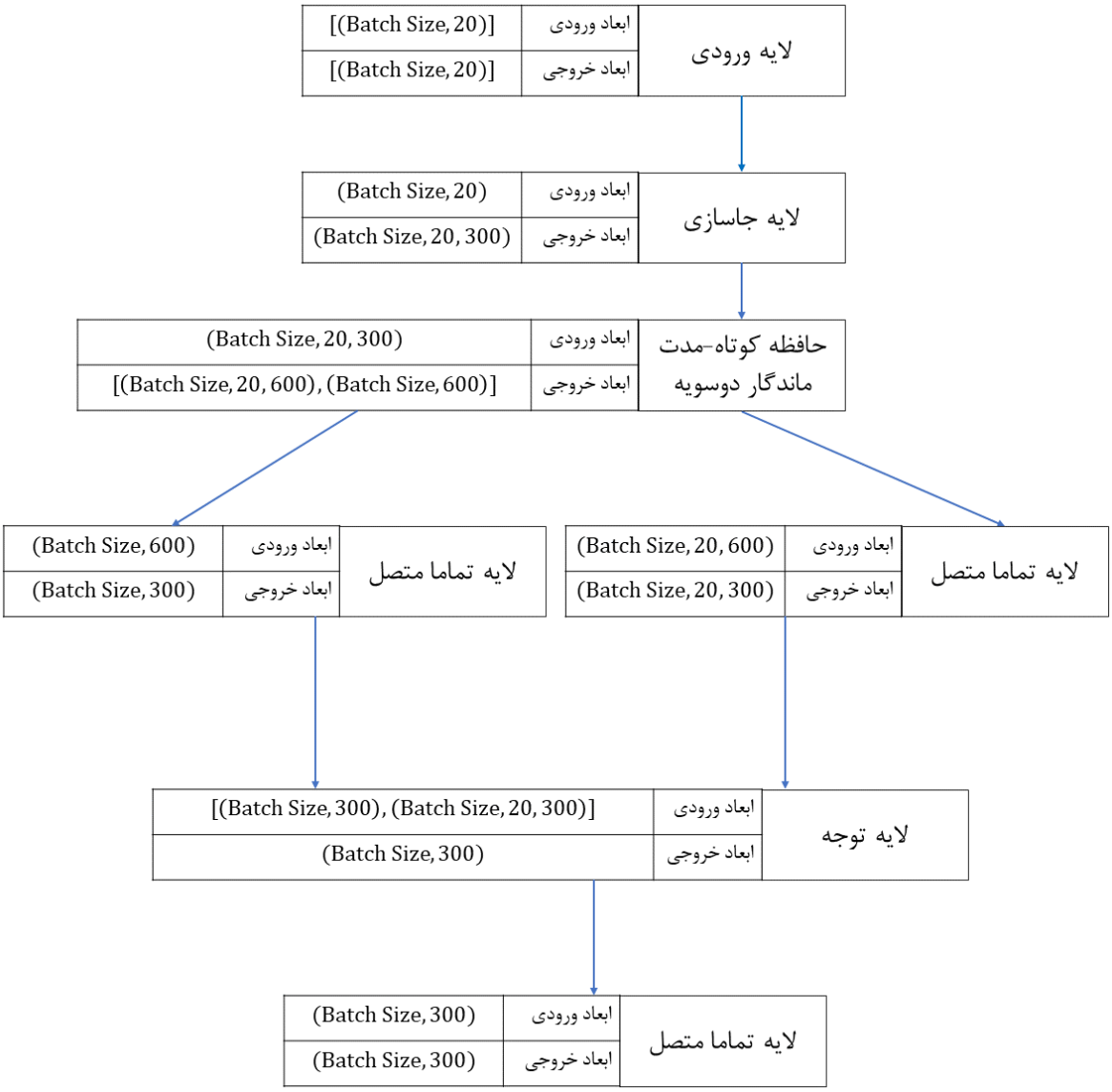
۴-۲-۴ مدل حافظه کوتاه-مدت ماندگار دوسویه با توجه

در این مدل، بردارهای بدست‌آمده از لایه جاسازی، وارد یک حافظه کوتاه-مدت ماندگار دوسویه می‌شوند. این لایه، یک بار ورودی را از ابتدا به انتها و بار دیگر آن را از انتها به ابتدا پردازش می‌کند. به همین دلیل، برای هر لحظه از زمان، ابتدا دو حالت مخفی بدست می‌آورد. در این مدل، این دو حالت مخفی به یکدیگر الحاق می‌شوند و برداری جدید بدست می‌آید که آن را به عنوان حالت مخفی نهایی در حافظه کوتاه-مدت ماندگار دوسویه در نظر می‌گیریم.

حافظه کوتاه-مدت ماندگار دوسویه، هم وضعیت مخفی در لحظه آخر و هم وضعیت مخفی تمام لحظات را تحویل می‌دهد. این خروجی‌ها وارد دو لایه تماما متصل جداگانه می‌شوند و بعد آن‌ها به نصف مقدار اولیه کاهش می‌یابد. خروجی‌های دو لایه تماما متصل، وارد لایه توجه می‌شوند و خروجی این لایه نیز توسط یک لایه تماما متصل دوباره به همان بُعد تصویر می‌گردد. این تصویر، خروجی نهایی مدل است. پارامترهای مدل در جدول ۴-۴ ذکر شده است و شکل ۴-۱۰ معماری آن را نشان می‌دهد.

اندازه دنباله ورودی	تعداد نورون‌های هر لایه تماما متصل
۲۰	۳۰۰
بعد بردارها در ماتریس جاسازی	تابع فعال‌سازی هر لایه تماما متصل
۳۰۰	تانژانت هایپربولیک
تعداد نورون‌های حافظه کوتاه-مدت ماندگار	
۳۰۰	

جدول ۴-۴: پارامترهای مدل حافظه کوتاه-مدت ماندگار دوسویه با توجه



شکل ۴-۱۰: معماری مدل حافظه کوتاه-مدت ماندگار دوسویه با توجه

۴-۳ جمع‌بندی

در این فصل، ابتدا مقدماتی از شبکه‌های عصبی مورد استفاده در این پژوهش بیان گردید. سپس با استفاده از مفاهیم شرح داده‌شده، مدل‌هایی برای حل مسئله مطرح گردید. در این مدل‌ها از لایه‌های تماما متصل و شبکه‌های عصبی بازگشتی از جمله حافظه کوتاه-مدت ماندگار (دوسویه) و لایه توجه استفاده شده است.

فصل پنجم
آزمایش‌ها

در این فصل، عملکرد روش‌های پیشنهادی برای تعیین نزدیک‌ترین کلمه به یک مفهوم بررسی می‌شود. ابتدا نحوه پیش‌پردازش داده‌ها به منظور آموزش شبکه‌های عصبی شرح داده می‌شود. سپس نحوه انتخاب بهترین پارامترها برای هر شبکه ذکر می‌گردد. خلاصه یافته‌های آزمایش‌ها در انتهای فصل ذکر شده است.

تمامی آزمایش‌ها با استفاده از زبان برنامه‌نویسی پایتون و روی بستر Google Collaboratory Pro انجام گرفته است که مشخصات سخت‌افزاری آن به شرح زیر می‌باشد.

CPU: Intel(R) Xeon(R) CPU @ 2.30GHz

GPU: Nvidia Tesla K80

RAM: 25.51 GB

Disk Space: 107.77 GB

همچنین از ماژول‌های زیر استفاده شده است.

- TensorFlow v2.2.0 برای پیاده‌سازی شبکه‌های عصبی
- Numpy v1.18 برای محاسبات ماتریسی
- Scipy v1.18.4 برای محاسبه سریع فواصل بردارها
- Scikit-learn v0.23.1 برای محاسبه امتیاز توافق [۱]
- Gensim v3.8.3 برای بارگذاری و استفاده از بازنمایی برداری کلمات
- Hazm v0.7.0 برای دسترسی به سرواژه [۲] کلمات

۵-۱ پیش‌پردازش

برای آموزش شبکه‌های عصبی، به کلمات و عبارات توصیفی نیاز است. در این پژوهش، کلمات و عبارات توصیفی از لغت‌نامه‌ها، صفحات ویکی‌پدیا و شبکه واژگانی فارسی (فارس نت) استخراج شده‌اند و در قالب چهارتایی‌های مرتب متشکل از عبارت توصیفی، کلمه، شناسه و منبع تحت عنوان «اطلاعات تفکیک‌شده کلمات» ذخیره شده‌اند. در این قسمت، مراحل پیش‌پردازش این داده‌ها شرح داده می‌شود.

[۱] Agreement score

[۲] Lemma

۵-۱-۱ حذف هر کلمه از توصیف خود

در برخی موارد، یک کلمه با استفاده از خود آن توصیف شده است. در این گونه موارد، کلمه را از توصیف خود حذف می‌کنیم؛ زیرا هر کلمه باید مستقل از خودش تعریف گردد.

۵-۱-۲ حذف ایست‌واژه‌ها [۳]

ایست‌واژه‌ها کلماتی هستند که به فراوانی در عبارات مختلف یافت می‌شوند و به مفهوم کل عبارت چیزی اضافه نمی‌کنند؛ به طور مثال، حروف ربط از ایست‌واژه‌ها محسوب می‌شوند. وجود این واژه‌ها در یک عبارت توصیفی، چیزی به مفهوم آن عبارت اضافه نمی‌کند. از طرفی، فراوانی این واژه‌ها در داده‌های آموزشی می‌تواند باعث شود مدل بیش از حد به آن‌ها بها دهد و الگوهایی را از داده‌ها کشف کند که در واقعیت وجود ندارند. به همین دلیل، این واژه‌ها نیز از عبارات توصیفی حذف شدند. همچنین، هر عبارتی که یکی از این واژه‌ها را توصیف می‌کند، حذف گردید.

۵-۱-۳ حذف کلمات و عبارات توصیفی کوتاه

در لغت‌نامه‌ها برای هر یک از حروف الفبای فارسی نیز توضیحاتی ذکر شده است؛ به طور مثال با مراجعه به مدخل حرف «د» می‌توانیم دریابیم «دهمین حرف از حروف الفبای فارسی» است. با توجه به مسئله این پژوهش، کاربر به دنبال یک کلمه است و لذا به عنوان خروجی انتظار دریافت یک حرف را ندارد. به همین دلیل در این مرحله، عبارات مربوط به توصیف حروف الفبا از مجموعه داده‌ها حذف شد.

همچنین، با بررسی کلمات دوحرفی موجود در لغت‌نامه‌ها مشخص شد این کلمات غالبا جزء یکی از دسته‌های زیر محسوب می‌شوند:

- ایست‌واژه‌ها (مثال: «هر»)
- اعداد (مثال: «سه»)
- پیشوندهای عربی (مثال: «ذو»)
- آواها (مثال: «آخ»)
- کلمات با کاربرد بسیار کم و یا منسوخ (مثال: «یی» به معنای «نام حرف یاء»)

[۳] Stopwords

از آنجایی که کاربر برای دست‌یابی به این گونه کلمات نیازی به یک سیستم هوشمند ندارد، این موارد از داده‌ها حذف شدند و بقیه موارد مربوط به کلمات دوحرفی در داده‌ها باقی ماندند. همچنین، هر عبارت توصیفی شامل فقط یک حرف یا کلمه‌ای دو حرفی که در دسته‌های بالا جای می‌گیرد، از مجموعه داده‌ها حذف شد؛ زیرا این گونه عبارت‌ها شامل اطلاعات کافی برای کمک به روند یادگیری مدل نیستند.

۵-۱-۴ واحدسازی [۴]

«اصل ترکیب معنایی» [۵] بیان می‌کند که معنای یک عبارت (فقط) تابعی معنای اجزاء آن و روش ترکیب آن‌ها می‌باشد [۳۵]. بر مبنای این اصل، هر عبارت توصیفی به مجموعه‌ای از واحدها تبدیل شد. جداسازی واحدها بر اساس رخداد کاراکتر فاصله انجام شده است.

۵-۱-۵ تقسیم داده‌ها

لغت‌نامه‌ها حاوی معانی مختلفی برای هر کلمه هستند. برای هر معنی، یک یا چند عبارت توصیفی وجود دارد. این عبارات، یک مفهوم را با استفاده از کلمات متفاوت توصیف می‌کنند. برای دسترسی به کلیه عبارات توصیفی حاصل از هر معنی کلمه‌ای مانند w که از منبع s استخراج شده است، کافیست ابتدا در بین چهارتایی‌های مرتب موجود در «اطلاعات تفکیک‌شده کلمات» مواردی را در نظر بگیریم که دومین عنصر آن‌ها برابر w و آخرین عنصر آن‌ها برابر s می‌باشد و حاصل جستجو را بر اساس سومین عنصر گروه‌بندی کنیم. فرض کنید یکی از معانی کلمه‌ای مانند w توسط c عبارت مانند $d_1, d_2, \ldots, d_c$ توصیف شده است. چهارتایی‌های متناظر این عبارات را به ۳ مجموعه افراز می‌کنیم؛ به طوری که تقریبا ۸۰ درصد آن‌ها در یک مجموعه، ۱۰ درصد در مجموعه دوم و ۱۰ درصد دیگر در مجموعه سوم قرار گیرد. سپس با حذف عناصر سوم و چهارم هر چهارتایی (شناسه و منابع)، هر کدام از اعضای این مجموعه‌ها به شکل یک دوتایی مرتب مانند (w, d_i) به ترتیب به داده‌های آموزشی، اعتبارسنجی و آزمایشی اضافه می‌شوند.

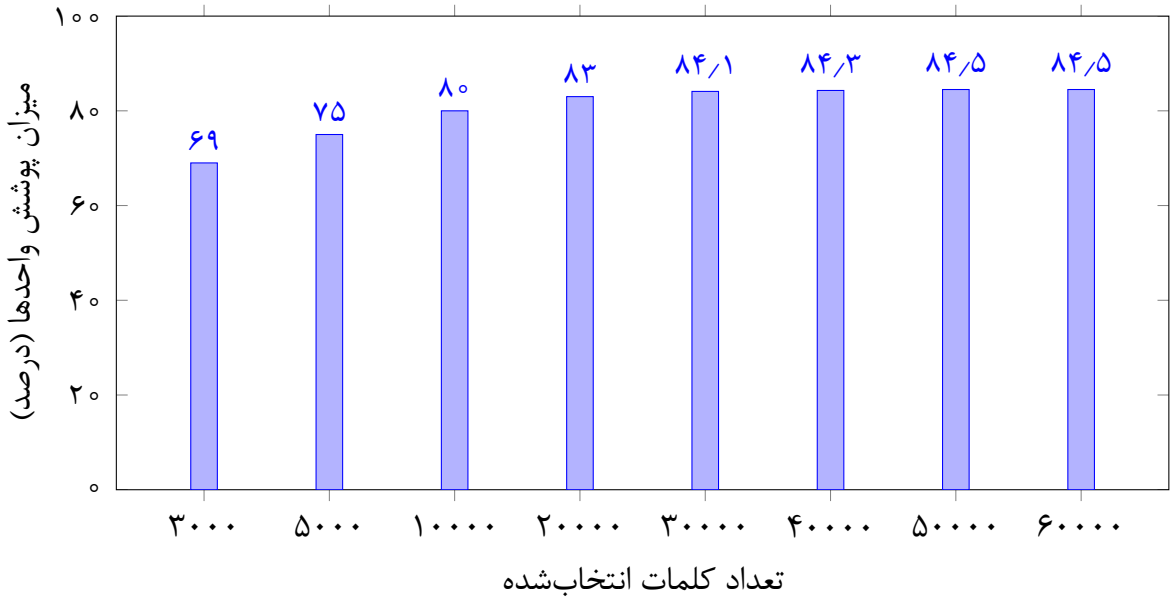
چهارتایی‌های بدست‌آمده از صفحات ویکی‌پدیا و فارس‌نت نیز مانند داده‌های لغت‌نامه‌ها به شکل دوتایی‌های مرتب شامل کلمه و عبارت توصیفی در می‌آیند و همگی به داده‌های آموزشی اضافه می‌شوند.

[۴] Tokenization

[۵] The principle of semantic compositionality

۵-۱-۶ نرمال‌سازی کلمات بر اساس رتبه

به دلیل محدودیت حافظه و زمان، در هر آزمایش بر اساس رتبه‌بندی کلمات تعدادی کلمه را که دارای بالاترین رتبه‌ها هستند، انتخاب می‌کنیم و تنها نمونه‌هایی را در نظر می‌گیریم که حاوی توصیف این کلمات هستند. برای بررسی میزان پوشش متون نوشتاری فارسی توسط این کلمات، از متون صفحات پیکره ویکی‌پدیای فارسی استفاده می‌کنیم. بر اساس رخداد کاراکتر فاصله، متون این صفحات را به تعدادی واحد تجزیه می‌کنیم. مشابه بخش ۳-۳-۲ هر واحد را نرمال می‌کنیم. ایست‌واژه‌ها را نیز حذف می‌کنیم. با استفاده از ماژول هضم، سرواژه هر واحد را جایگزین آن می‌کنیم. سپس میزان پوشش واحدهای نهایی توسط کلمات انتخابی را محاسبه می‌نماییم. شکل ۵-۱ میزان پوشش این واحدها را به ازای انتخاب تعداد مختلفی از کلمات نشان می‌دهد.



شکل ۵-۱: میزان پوشش پیکره ویکی‌پدیای فارسی توسط کلمات انتخاب‌شده بر اساس رتبه‌بندی

با توجه به اینکه با افزایش تعداد کلمات انتخاب‌شده از ۲۰۰۰۰ به مقادیر بیشتر، به ازای هر ۱۰۰۰۰ کلمه حداکثر به اندازه ۱ درصد تغییر در میزان پوشش پیکره ایجاد می‌شود، برای آزمایش‌های این پژوهش مقادیر ۳۰۰۰، ۵۰۰۰، ۱۰۰۰۰ و ۲۰۰۰۰ را برای تعداد کلمات انتخابی در نظر می‌گیریم.

۵-۱-۷ داده‌افزایی [۶]

در برخی آزمایش‌ها، با کمبود داده‌های آموزشی مواجه هستیم. به همین دلیل، تعدادی از داده‌های آموزشی موجود را انتخاب می‌کنیم. در هر کدام از این داده‌ها، در عبارات توصیفی به دنبال کلماتی می‌گردیم که حداقل یک مترادف برای آن‌ها موجود باشد. اگر در یک عبارت توصیفی، کلمه‌ای را با مترادف آن جایگزین کنیم، معمولا ساختار جمله حفظ می‌شود و یک داده آموزشی جدید بدست می‌آید. در این پژوهش، برای داده‌افزایی از روی هر عبارت توصیفی، k عبارت توصیفی دیگر بدست آمده است که هر کدام در w کلمه با عبارت توصیفی اولیه تفاوت دارند. این عبارات توصیفی جدید، به مجموعه داده‌های آموزشی افزوده می‌شود.

۵-۱-۸ کدگذاری [۷] کلمات و تشکیل ماتریس مقایسه

برای هر کلمه باقی‌مانده که توصیفی از آن وجود دارد، یک شناسه در نظر گرفته شد. با استفاده از یک مدل برداری، برای هر کلمه یک بردار بدست آمد. از روی هم قرار گرفتن این بردارها، یک ماتریس تشکیل شد که آن را ماتریس مقایسه می‌نامیم. سطر i-ام این ماتریس، دربرگیرنده بردار مربوط به کلمه‌ای است که شناسه آن i می‌باشد. از این ماتریس برای مقایسه خروجی شبکه‌های عصبی با بازنمایی برداری کلیه کلمات استفاده خواهد شد. نحوه این مقایسه به طور دقیق در بخش ۵-۲-۳ توضیح داده شده است.

۵-۱-۹ نرمال‌سازی عبارات توصیفی در سطح واحد

در هر آزمایش، برای هر واحد در کل عبارات توصیفی موجود در داده‌های آموزشی، تعداد رخداد آن محاسبه شد. بر این اساس، واحدها از نظر میزان احتمال رخداد، رتبه‌بندی شدند. در نهایت حداکثر ۱۰۰۰۰۰ واحد با رتبه بالاتر انتخاب شد. سپس، هر عبارت توصیفی که شامل واحدی به جز این واحدها بود، از مجموعه داده‌های آموزشی، اعتبارسنجی و آزمایشی حذف گردید؛ در این گونه عبارت‌ها از کلماتی استفاده شده است که به ندرت در توصیف یک کلمه دیگر به کار رفته‌اند. وجود این گونه کلمات با کاربرد کم، می‌تواند باعث شود مدل رخداد آن‌ها را به عنوان یک الگو بشناسد. با حذف عبارات شامل این کلمات، از شناسایی این الگوهای غیر واقعی جلوگیری به عمل می‌آید.

[۶] Data augmentation

[۷] Encoding

۵-۱-۱۰ کدگذاری واحدها

به هر واحد، یک شناسه نسبت داده شد. همچنین، دو واحد جدید با نمادهای PAD برای نمایش یک واحد خالی و UNK برای نمایش هر واحد ناشناخته در نظر گرفته شد. شناسه واحدهای PAD و UNK به ترتیب برابر ۰ و ۱ تعیین شد. بدین ترتیب یک کدگذاری برای واحدها حاصل شد. از این کدگذاری در تبدیل عبارت ورودی کاربر به یک لیست از اعداد استفاده می‌شود. این لیست از اعداد به عنوان ورودی به مدل‌های پیشنهادی داده خواهد شد.

۵-۱-۱۱ تشکیل ماتریس جاسازی

برای هر واحد باقی‌مانده در عبارات توصیفی، یک سطر متناظر شناسه آن در ماتریس جاسازی در نظر گرفته شد؛ به طوری که سطر i-ام این ماتریس، متناظر واحد با شناسه i باشد. مقدار اولیه عناصر هر سطر، با بردار کلمه متناظر آن در یک مدل برداری، مشخص شد. از این ماتریس در لایه جاسازی استفاده خواهد شد.

۵-۱-۱۲ اصلاح طول عبارات توصیفی

مدل‌های پیشنهادی، لیستی به طول ۲۰ را می‌پذیرند. در همین راستا، هر عبارت توصیفی با طول بیش از ۲۰ واحد، کوتاه شده و تنها ۲۰ واحد اول آن در نظر گرفته شد. همچنین، با افزودن واحدهای PAD به هر عبارت با طول کمتر از ۲۰ واحد، طول آن‌ها به ۲۰ واحد رسانده شد. به این ترتیب، تمامی عبارات توصیفی دقیقا شامل ۲۰ واحد شدند.

۵-۱-۱۳ حذف داده‌های تکراری

در برخی موارد، توصیف یک کلمه در یک لغت‌نامه، از لغت‌نامه‌ای دیگر اقتباس شده است. به همین دلیل، داده‌های تکراری در لغت‌نامه‌ها یافت می‌شود. فرض کنید دوتایی مرتب (w, d) بیش از یک بار در داده‌ها وجود داشته باشد. در این صورت مدل تلاش بیشتری را در جهت یادگیری این دوتایی مرتب به عمل خواهد آورد. به بیان دیگر، مدل اهمیت بیشتری به این دوتایی مرتب خواهد داد؛ در حالی که در واقعیت این دوتایی مرتب اهمیت بیشتری نسبت به بقیه دوتایی‌های مرتب ندارد. لذا تکرار رخداد آن در داده‌ها می‌تواند روند یادگیری مدل را مختل نماید. از همین رو، از هر دوتایی مرتب به شکل کلمه و

عبارت توصیفی که بیش از یک بار در داده‌ها موجود است، یک نمونه در داده‌های آموزشی باقی ماند و بقیه موارد حذف شدند.

۵-۱-۱۴ تشکیل نمونه‌های آموزشی، اعتبارسنجی و آزمایشی

ورودی هر مدل پیشنهادی، یک لیست از اعداد است؛ در حالی که داده‌های ما از جنس کلمات هستند. با استفاده از کدگذاری واحدهای موجود در عبارات توصیفی، هر کدام از آن‌ها را به یک لیست با طول ۲۰ از اعداد صحیح می‌نگاریم. در ادامه، مثالی از این نگاشت آورده شده است.

نمونه ۵-۱-۱. نگاشت یک لیست از واحدها به لیستی از اعداد صحیح نامنفی

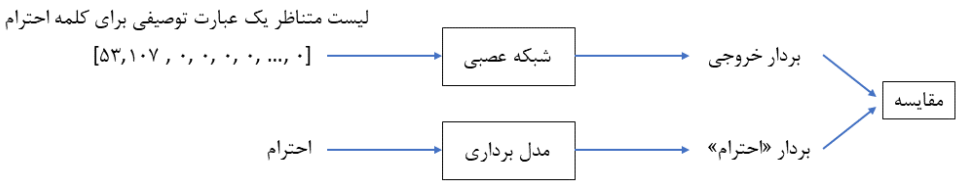
کلمه: احترام

عبارت توصیفی: ارج نهادن

لیست اولیه: [ارج، نهادن، PAD، PAD، PAD، PAD، ...، PAD]

حاصل: [۵۳, ۱۰۷, ۰, ۰, ۰, ۰, ..., ۰]

با این نگاشت، مدل‌های پیشنهادی می‌توانند عبارات توصیفی را در قالب لیست‌هایی از اعداد بپذیرند. به بیان دیگر، این لیست‌ها نمایشی از عبارات توصیفی هستند که هر کدام از آن‌ها، متناظر یک کلمه می‌باشند. برای مقایسه خروجی مدل با این کلمه، باید کلمه را به شکل دیگری نمایش دهیم. از آنجایی که خروجی هر مدل یک بردار است، کلمه متناظر هر لیست را به وسیله برداری از همان بُعد نمایش می‌دهیم. این بردار از طریق یک مدل برداری بدست می‌آید. عملکرد مدل زمانی قابل قبول خواهد بود که بردار خروجی آن، نزدیک به بردار متناظر کلمه موردنظر باشد. در نمونه بالا انتظار داریم خروجی مدل، برداری نزدیک به بردار کلمه «احترام» باشد. شکل ۵-۲ روند مقایسه خروجی مدل و خروجی مورد انتظار را نشان می‌دهد.



شکل ۵-۲: مقایسه خروجی مدل و خروجی مورد انتظار

بدین ترتیب، از هر دوتایی مرتب متشکل از کلمه و عبارت توصیفی، یک لیست و یک بردار بدست

می‌آید. با این روش، دوتایی‌های مرتب موجود در داده‌های آموزشی، اعتبارسنجی و آزمایشی را به نمونه‌هایی متشکل از لیست و بردار تبدیل می‌کنیم. لیست موجود در هر نمونه به عنوان ورودی وارد شبکه عصبی می‌شود و بردار آن نمونه، با خروجی شبکه مقایسه می‌گردد.

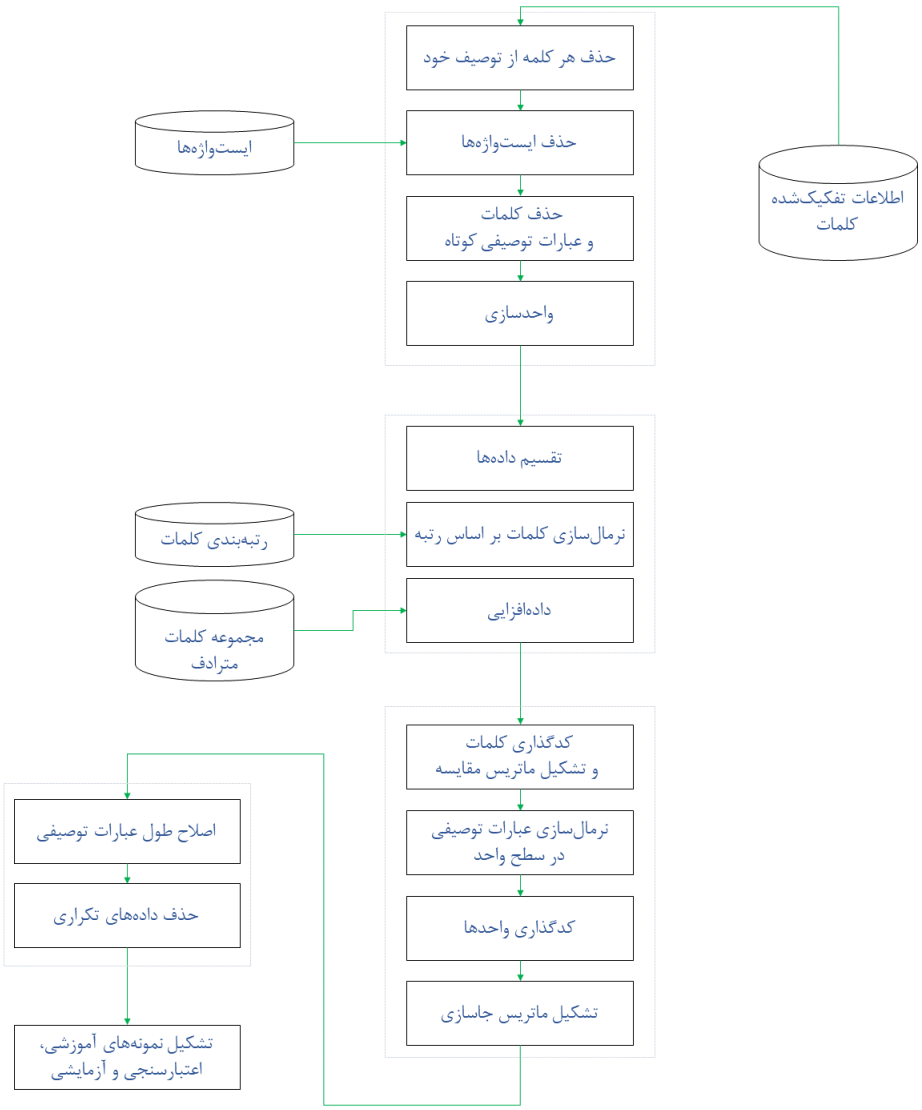
علاوه بر آن، برای نمونه‌های آموزشی، اعتبارسنجی و آزمایشی، سه لیست جدا تشکیل می‌شود که هر کدام متشکل از کلمات متناظر نمونه‌ها است. این لیست‌ها را به ترتیب کلمات صحیح آموزشی، اعتبارسنجی و آزمایشی می‌نامیم.

در نهایت با توجه به روند تقسیم داده‌ها و تشکیل نمونه‌ها، تعداد نمونه‌های آموزشی، اعتبارسنجی و آزمایشی با توجه به نرمال‌سازی کلمات بر اساس رتبه، در جدول ۵-۱ ذکر شده است.

پایین‌ترین رتبه کلمات	تعداد نمونه‌های آموزشی	تعداد نمونه‌های اعتبارسنجی	تعداد نمونه‌های آزمایشی
۳۰۰۰	۱۰۱۷۹۶	۹۳۵۲	۸۰۴۹
۵۰۰۰	۱۴۶۲۰۳	۱۴۶۵۷	۱۲۲۱۰
۱۰۰۰۰	۲۳۴۶۵۰	۲۴۹۰۲	۲۰۳۴۴
۲۰۰۰۰	۳۳۷۹۸۴	۳۷۳۷۸	۳۰۶۴۶

جدول ۵-۱: آمار تقسیم نمونه‌ها با توجه به انتخاب کلمات بر اساس رتبه

شکل ۵-۳ تمامی مراحل پیش‌پردازش را به صورت اجمالی شرح می‌دهد.



شکل ۵-۳: مراحل پیش‌پردازش داده‌ها

۵-۲ معیارهای ارزیابی

فرض کنید لیست موجود در هر نمونه را با L و بردار متناظر آن را با V نمایش دهیم. در نظر بگیرید M یک مدل باشد و خروجی آن روی L را با $M(L)$ نمایش دهیم. در ادامه بر پایه این نمادگذاری، معیارهایی را برای ارزیابی عملکرد مدل روی هر نمونه ارائه می‌کنیم.

۵-۲-۱ شباهت و خطای کسینوسی

شباهت کسینوسی دو بردار مانند V_1 و V_2 در فضای $\mathbb{R}^m$ از طریق رابطه زیر بدست می‌آید.

$$cos(V_1, V_2) = \frac{V_1 \cdot V_2}{|V_1||V_2|} \quad (۱-۵)$$

در رابطه فوق، $\cdot$ نماد ضرب نقطه‌ای و $||$ نشان‌دهنده نرم اقلیدسی می‌باشد. بر همین اساس، شباهت کسینوسی دو بردار $M(L)$ و V به عنوان یک معیار از عملکرد مدل در نظر گرفته می‌شود. هر چه این شباهت بیشتر باشد، عملکرد مدل روی آن نمونه بهتر بوده است. بر همین اساس، می‌توان خطای کسینوسی مدل روی نمونه ($cosine\ loss(M(L), V)$) را نیز به شکل زیر تعریف نمود.

$$cosine\ loss(M(L), V) = -cos(M(L), V) \quad (۲-۵)$$

۵-۲-۲ خطای رتبه‌ای

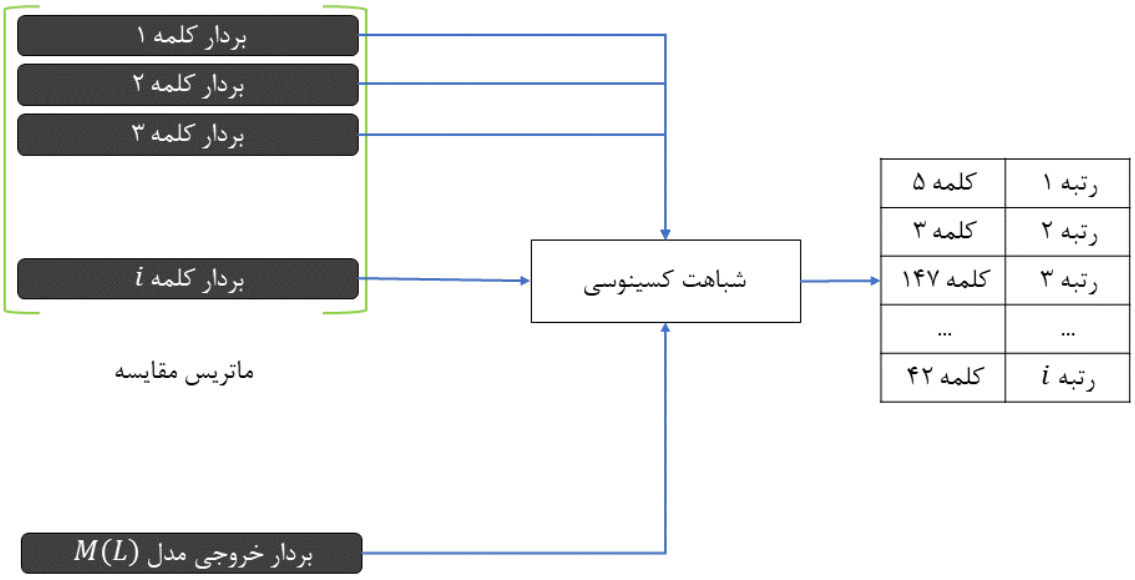
برای هر نمونه مانند (L, V) می‌دانیم V بردار متناظر یک کلمه است. فرض کنید V' بردار کلمه‌ای دیگر باشد که به صورت تصادفی از ماتریس مقایسه انتخاب شده است. مطلوب آن است که خروجی مدل یعنی $M(L)$ به V نزدیک و از V' دور باشد. به بیان دیگر هر چه $cos(M(L), V') - cos(M(L), V)$ کاهش یابد، عملکرد مدل بهتر می‌شود. بر این اساس، خطای رتبه‌ای از طریق رابطه زیر محاسبه می‌گردد.

$$rank\ loss(M(L), V) = max(0, m + cos(M(L), V') - cos(M(L), V)) \quad (۳-۵)$$

در رابطه فوق، m یک عدد ثابت است که میزان فاصله مطلوب میان دو مقدار $cos(M(L), V')$ و $cos(M(L), V)$ را نشان می‌دهد. در این پژوهش، m برابر ۱ در نظر گرفته شده است. برای اینکه تابع مشتق‌پذیر باشد، پس از محاسبه اختلاف دو شباهت کسینوسی و افزودن m به آن، اگر حاصل کمتر از صفر شد، خطای نمونه را برابر ۰ در نظر می‌گیریم.

۵-۲-۳ دقت در k

خروجی مدل به ازای هر لیست ورودی مانند L، یک بردار مانند $M(L)$ است. اگر شباهت کسینوسی $M(L)$ با هر سطر ماتریس مقایسه را محاسبه کنیم، در واقع شباهت خروجی مدل با بردار هر کلمه را سنجیده‌ایم. از آنجایی که بردارهای کلمات بر اساس کدگذاری آن‌ها در ماتریس مقایسه قرار گرفته‌اند، اگر مقادیر شباهت‌های کسینوسی را از بزرگ به کوچک مرتب کنیم، به یک رتبه‌بندی از کلمات مانند R می‌رسیم.



شکل ۵-۴: نحوه استفاده از ماتریس مقایسه

از طرفی، با استفاده از لیست کلمات صحیح، می‌توان به کلمه‌ای که لیست L متناظر آن است، دست یافت. اگر این کلمه w باشد، مکان آن در رتبه‌بندی R را رتبه کلمه صحیح می‌نامیم و با $index(w, R)$ نشان می‌دهیم. همچنین، اگر $synset(w) = \{s_1(w), s_2(w), \ldots, s_n(w)\}$ مجموعه کلمات مترادف w باشد، مکان هر یک از آن‌ها در R را محاسبه می‌کنیم و بالاترین رتبه را با $index(synset(w), R)$ نشان می‌دهیم. در واقع داریم:

$$index(synset(w), R) = min\{index(s_i(w), R) : i \in \{1, 2, \ldots, n\}\} \qquad (۵-۴)$$

پس از محاسبه $index(w, R)$ و $index(synset(w), R)$ برای کلیه نمونه‌های مورد بررسی، هر کدام از این دو مقدار را در یک لیست جداگانه قرار می‌دهیم. فرض کنید مقادیر $index(w, R)$ در لیست

L_{exact} و مقادیر $index(synset(w), R)$ در لیست L_{syn} قرار گیرند.

دقت در k را برای هر لیست مانند L از اعداد صحیح نامنفی با نماد $Acc@k$ نشان می‌دهیم و به شکل زیر تعریف می‌نماییم.

$$Acc@k := \frac{|\{x \in L : x \leq k\}|}{|L|} \tag{۵-۵}$$

در این پژوهش، از نمونه‌های آموزشی و آزمایشی، به صورت تصادفی دو زیرمجموعه کوچکتر (شامل ۵۰۰ نمونه) انتخاب می‌کنیم و دقت در ۱۰ و ۱۰۰ را برای هر کدام از آن‌ها محاسبه می‌نماییم. بدین ترتیب، مقادیر $Acc@۱۰$ و $Acc@۱۰۰$ را برای زیرمجموعه‌های آموزشی و آزمایشی، محاسبه و گزارش می‌نماییم. این مقادیر را برای L_{exact} با نمادهای $Acc_{exact}@۱۰$ و $Acc_{exact}@۱۰۰$ و برای L_{syn} با نمادهای $Acc_{syn}@۱۰$ و $Acc_{syn}@۱۰۰$ نشان می‌دهیم.

۵-۲-۴ شاخص نظرخواهی میانگین [۸]

شاخص نظرخواهی میانگین، یک معیار ارزیابی است که با نظرخواهی از انسان‌های داوطلب بدست می‌آید. برای محاسبه این مقدار، ابتدا پرسش‌هایی در اختیار داوطلبان قرار می‌گیرد. هر داوطلب موظف است در پاسخ به هر پرسش امتیازی صحیح از ۱ تا ۵ را اعلام کند. در نهایت میانگین امتیازات اعلام‌شده توسط داوطلبان، شاخص نظرخواهی میانگین را مشخص می‌کند.

فرض کنید کلمه w به وسیله عبارت d در یک لغت‌نامه توصیف شده باشد و L_d لیستی باشد که از پیش‌پردازش d بدست آمده است. اگر M مدلی باشد که L_d را به عنوان ورودی می‌پذیرد، با محاسبه شباهت کسینوسی میان $M(L_d)$ و سطرهای ماتریس مقایسه، می‌توان به یک رتبه‌بندی از کلمات دست یافت. می‌توان کلمه دارای رتبه r در این رتبه‌بندی را به عنوان پیشنهاد r-ام مدل به کاربر در نظر گرفت.

در این پژوهش، قصد مقایسه سه پیشنهاد اول مدل با کلمه w را داریم. برای هر نمونه، یک کلمه صحیح، ۳ پیشنهاد مدل و یک عبارت توصیفی وجود دارد. می‌خواهیم مشخص کنیم هر کدام از این ۴ کلمه چقدر مفهوم آن عبارت توصیفی را منتقل می‌کنند. برای هر لغت‌نامه، ۴۰ عبارت توصیفی به همراه کلمه صحیح و ۳ پیشنهاد مدل را در نظر گرفتیم. بر این اساس ۴ پرسشنامه طراحی شد. پرسش‌نامه اول، میزان انطباق هر کلمه صحیح (موجود در لغت‌نامه) با عبارت توصیفی متناظر آن را می‌سنجد. پرسش‌نامه دوم، میزان انطباق پیشنهاد اول مدل با عبارت توصیفی متناظر آن را بررسی می‌کند. پرسش‌نامه‌های

[۸] Mean opinion score

سوم و چهارم نیز به ترتیب همین مورد را درباره پیشنهادهای دوم و سوم مدل بررسی می‌کنند. نمونه ۵-۲-۱ مثالی از نحوه انتخاب سوالات پرسش‌نامه‌ها را نشان می‌دهد.

نمونه ۵-۲-۱. بدست‌آوردن پرسش‌ها از روی نمونه‌های آزمایشی

نمونه آزمایشی:

([۵۳, ۱۰۷, ۰, ۰, ۰, ۰, . . . , ۰], [۰٫۳۵, ۱٫۶۷, −۰٫۶۶, . . . , ۰٫۸۹])

کلمه متناظر نمونه آزمایشی در لیست کلمات صحیح: احترام

عبارت توصیفی متناظر نمونه: ارج‌نهادن

از کاربر خواسته می‌شود مشخص کند «احترام»، به چه میزان مفهوم «ارج‌نهادن» را منتقل می‌کند. این سوال در پرسش‌نامه اول درج می‌شود.

پیشنهاد اول مدل: تکریم

از کاربر خواسته می‌شود مشخص کند «تکریم»، به چه میزان مفهوم «ارج‌نهادن» را منتقل می‌کند. این سوال در پرسش‌نامه دوم درج می‌شود.

پیشنهاد دوم مدل: صبا

از کاربر خواسته می‌شود مشخص کند «صبا»، به چه میزان مفهوم «ارج‌نهادن» را منتقل می‌کند. این سوال در پرسش‌نامه سوم درج می‌شود.

پیشنهاد سوم مدل: بزرگداشت

از کاربر خواسته می‌شود مشخص کند «بزرگداشت»، به چه میزان مفهوم «ارج‌نهادن» را منتقل می‌کند. این سوال در پرسش‌نامه چهارم درج می‌شود.

در این پژوهش، ۲۱ نفر از دانشجویان مقطع کارشناسی رشته زبان و ادبیات فارسی و ۸ نفر از دانشجویان مقطع کارشناسی ارشد رشته زبان‌شناسی همگانی به عنوان شرکت‌کننده به پرسش‌نامه‌ها پاسخ داده‌اند. پس از دریافت پاسخ‌های یک پرسش‌نامه از کلیه شرکت‌کنندگان، لازم است میزان اعتبار پاسخ‌های هر شرکت‌کننده بررسی شود. پس از این بررسی می‌توان، پاسخ‌های کم‌اعتبار را حذف و در نهایت شاخص نظرخواهی میانگین را برای پاسخ‌های باقی‌مانده محاسبه نمود. در ادامه نحوه شناسایی پاسخ‌های کم‌اعتبار شرح داده می‌شود.

شناسایی و حذف پاسخ‌های کم‌اعتبار با استفاده از امتیاز توافق [۹]

فرض کنید q سوال در یک پرسشنامه موجود باشد و n نفر به سوالات پرسشنامه پاسخ داده باشند. همچنین، در نظر بگیرید که هر نفر بتواند برای هر سوال یکی از پاسخ‌های $a_۱, a_۲, \ldots, a_c$ را اعلام کند. برای هر دو شرکت‌کننده مانند h و h' ، اگر f_{ij} تعداد سوالاتی را نشان دهد که h و h' به ترتیب پاسخ a_i و a_j را برایشان اعلام کرده‌اند، توافق نسبی مشاهده‌شده [۱۰] برای h و h' از طریق رابطه ۵-۶ بدست می‌آید.

$$p_o = \frac{۱}{n} \sum_{i=۱}^{c} f_{ii} \tag{۵-۶}$$

همچنین، توافق مورد انتظار [۱۱] از طریق رابطه ۵-۷ محاسبه می‌گردد.

$$p_e = \frac{۱}{n^۲} \sum_{i=۱}^{c} (\sum_{j=۱}^{c} f_{ij} \sum_{j=۱}^{c} f_{ji}) \tag{۵-۷}$$

پس از محاسبه این دو مقدار، برای هر دو شرکت‌کننده مانند h و h' امتیاز توافق با نماد κ نشان داده و از طریق رابطه ۵-۸ محاسبه می‌شود [۴۸].

$$\kappa = \frac{p_o - p_e}{۱ - p_e} \tag{۵-۸}$$

اولین بار Cohen [۹] این امتیاز توافق را معرفی نمود. با محاسبه مقدار κ برای هر دو شرکت‌کننده، میزان شباهت پاسخ‌هایشان مشخص می‌گردد؛ اما این نحوه محاسبه امتیاز توافق قادر به لحاظ‌کردن سطوح مختلف عدم توافق نمی‌باشد. برای مثال فرض کنید سه شرکت‌کننده مانند h ، h' و h'' برای یک سوال به ترتیب پاسخ‌های ۱، ۲ و ۳ را اعلام کنند. با توجه به نحوه تعریف توافق مشاهده‌شده و توافق مورد انتظار، میزان اختلاف نظر میان h و h' برای این سوال با میزان اختلاف نظر میان h و h'' برابر در نظر گرفته می‌شود؛ در حالی که اختلاف نظر میان h و h'' بیشتر از h و h' است. برای در نظر گرفتن سطوح مختلف عدم توافق، Cohen [۱۰] نسخه وزن‌داری از امتیاز توافق را معرفی نمود. در این نسخه، برای هر دو پاسخ مانند a_i و a_j یک وزن مانند w_{ij} در نظر گرفته می‌شود. پس از تعیین وزن‌ها، توافق

[۹] Agreement score

[۱۰] Observed proportional agreement

[۱۱] Expected agreement

مشاهده‌شده و توافق مورد انتظار از طریق روابط ۵-۹ و ۵-۱۰ محاسبه می‌گردد.

$$p_o = \frac{۱}{n}\sum_{i=۱}^{c}\sum_{j=۱}^{c} w_{ij} f_{ij} \tag{۵-۹}$$

$$p_e = \frac{۱}{n^۲}\sum_{i=۱}^{c}\sum_{j=۱}^{c} w_{ij}(\sum_{k=۱}^{c} f_{ik} \sum_{k=۱}^{c} f_{kj}) \tag{۵-۱۰}$$

با این تعاریف جدید از توافق مشاهده‌شده و توافق مورد انتظار، می‌توان امتیاز توافق وزن‌دار را مانند گذشته از رابطه ۵-۸ محاسبه نمود.

در این پژوهش برای تعیین وزن‌ها از روش خطی استفاده کرده‌ایم. در این روش که توسط Cicchetti و Allison [۸] پیشنهاد شد، داریم:

$$w_{ij} = ۱ - \frac{|i-j|}{(c-۱)} \tag{۵-۱۱}$$

امتیاز توافق، میزان شباهت پاسخ‌های دو شرکت‌کننده را تعیین می‌کند؛ اما با استفاده از آن می‌توان میزان شباهت پاسخ‌های هر شرکت‌کننده با دیگران را نیز مشخص نمود. فرض کنید $h_۱, h_۲, \ldots, h_n$ کلیه شرکت‌کنندگان در فرآیند پاسخ‌دهی به سوالات یک پرسش‌نامه باشند و $\kappa(h_r, h_t)$ میزان توافق وزن‌دار (با وزن‌دهی خطی) برای دو شرکت‌کننده h_r و h_t را نشان دهد. در این صورت امتیاز توافق هر شرکت‌کننده مانند h_r با بقیه شرکت‌کنندگان را با نماد $\kappa_{total}(h_r)$ نشان داده و از طریق رابطه ۵-۱۲ بدست می‌آوریم.

$$\kappa_{total}(h_r) = \frac{\sum_{t=۱}^{n} \kappa(h_r, h_t)}{n} \tag{۵-۱۲}$$

پس از محاسبه مقدار κ_{total} برای هر شرکت‌کننده، با در نظر گرفتن یک آستانه [۱۲] می‌توان پاسخ‌های شرکت‌کننده‌هایی را که امتیاز توافق آن‌ها با بقیه کمتر از آن آستانه بوده است، حذف نمود. Landis و Koch [۲۵] برای تفسیر مقدار امتیاز توافق، بازه ۰ تا ۱ را به محدوده‌هایی تقسیم نمودند که در جدول ۵-۲ مشاهده می‌شود.

[۱۲] Threshold

درجه توافق	امتیاز توافق
ضعیف	۰.۰۰
کم	۰.۰۰-۰.۲۰
منصفانه	۰.۲۱-۰.۴۰
متوسط	۰.۴۱-۰.۶۰
قابل توجه	۰.۶۱-۰.۸۰
عالی	۰.۸۱-۱.۰۰

جدول ۵-۲: محدوده‌هایی برای تفسیر امتیاز توافق [۲۵]

با توجه به این محدوده‌ها و بررسی داده‌ها، در این پژوهش آستانه برابر ۰.۴ در نظر گرفته شد و پاسخ‌های هر شرکت‌کننده با مقدار κ_{total} کمتر از این آستانه حذف گردید.

۵-۳ انتخاب بهترین مدل

برای انتخاب بهترین مدل به منظور حل مسئله، مدل‌های پیشنهادی با تنظیمات متفاوتی آموزش داده شده‌اند. هدف این قسمت، مقایسه عملکرد این مدل‌ها پس از انتخاب بهترین تنظیمات آموزشی برای هر کدام از آن‌ها می‌باشد.

جدول ۵-۳ ، تنظیمات آموزشی برای مدل‌های پیشنهادی را نشان می‌دهد.

تنظیمات آموزشی	موارد بررسی‌شده
مدل برداری برای مقداردهی اولیه ماتریس جاسازی	FT, HAM-W2V, WIKI-W2V
امکان بروزرسانی مولفه‌ها در ماتریس جاسازی	بله/خیر
مقداردهی اولیه وزن‌های هر لایه	glorot uniform
اندازه دسته[۱۳]	۱۶، ۳۲، ۶۴، ۱۲۸، ۲۵۶
نرخ یادگیری	1.0, 0.1, 0.01, 0.001
تابع خطا	خطای کسینوسی - خطای رتبه‌ای
تعداد پیشنهادهای مدل	۳۰۰۰، ۵۰۰۰، ۱۰۰۰۰، ۲۰۰۰۰
استفاده از روش داده‌افزایی	بله/خیر
نوع توجه	جمع‌شونده - ضرب نقطه‌ای مقیاس‌پذیر
روش بهینه‌سازی	SGD, Adam, Adadelta
روش پیشگیری از بیش‌برازش	توقف زودهنگام، تنزل وزن، حذف تصادفی

جدول ۵-۳: تنظیمات آموزشی مورد بررسی برای هر مدل

برای هر مدل، ابتدا تعداد پیشنهادها ۳۰۰۰ فرض شده و بهترین تنظیمات با این فرض بدست آمده است. سپس برای آن تنظیمات خاص، تعداد پیشنهادها به ۵۰۰۰، ۱۰۰۰۰ و ۲۰۰۰۰ تغییر داده شده و نتایج ارزیابی مدل برای هر کدام از این موارد گزارش شده است.

[۱۳] Batch

۵-۳-۱ مدل سبد کلمات

جدول ۵-۴ بهترین تنظیمات مدل سبد کلمات را نشان می‌دهد.

اندازه دسته	مدل برداری	بروزرسانی ماتریس جاسازی	داده‌افزایی
۱۶	FT	بله	خیر
تابع خطا	روش بهینه‌سازی	نرخ یادگیری اولیه	پیشگیری از بیش‌برازش
خطای کسینوسی	Adadelta	۱.۰	توقف زودهنگام

جدول ۵-۴: بهترین تنظیمات برای مدل سبد کلمات

در جدول‌های ۵-۵ و ۵-۶ ، نتایج ارزیابی بهترین مدل روی داده‌های آموزشی و آزمایشی با توجه به سه معیار خطای کسینوسی، دقت در ۱۰ و ۱۰۰ آورده شده است.

۳۰۰۰ کلمه		۵۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۵۴		۰.۵۵	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۰۵	۰.۲۷۵	۰.۱	۰.۲۷۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۳	۰.۶	۰.۳۷۵	۰.۶۵
۱۰۰۰۰ کلمه		۲۰۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۵۶		۰.۵۷	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۱۲۵	۰.۲۲۵	۰.۱۵	۰.۲۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۴	۰.۵۷۵	۰.۴۲۵	۰.۵۷۵

جدول ۵-۵: ارزیابی مدل سبد کلمات روی داده‌های آموزشی

۳۰۰۰ کلمه		۵۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۶۳		۰.۶۳	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۰۷۵	۰.۱۲۵	۰.۰۲۵	۰.۱۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۲	۰.۵۲۵	۰.۳	۰.۶
۱۰۰۰۰ کلمه		۲۰۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۶۳		۰.۶۳	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۰۵	۰.۱۲۵	۰.۱۲۵	۰.۲۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۳۷۵	۰.۶	۰.۴۵	۰.۶۷۵

جدول ۵-۶: ارزیابی مدل سبد کلمات روی داده‌های آزمایشی

جدول ۵-۷ نتایج ارزیابی بهترین مدل بر اساس شاخص نظرخواهی میانگین به تفکیک تعداد پیشنهادها را نشان می‌دهد.

۳۰۰۰ کلمه				۵۰۰۰ کلمه			
عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۲	۱.۷	۱.۶	۱.۸	۳.۸	۲.۳	۲.۳	۱.۸
معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲.۹	۱.۴	۱.۲	۱.۲	۳.۴	۱.۷	۱.۵	۲
دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۱.۶	۱.۶	۱.۳	۱.۶	۲.۴	۱.۹	۱.۷	۱.۶
۱۰۰۰۰ کلمه				۲۰۰۰۰ کلمه			
عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۴	۲.۴	۲.۱	۲.۲	۳.۱	۲.۴	۲.۶	۲
معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۳	۲.۱	۱.۹	۱.۷	۲.۹	۲.۶	۲	۲.۱
دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲.۴	۲.۲	۱.۶	۲	۱.۸	۲.۲	۱.۹	۱.۶

جدول ۵-۷: ارزیابی شاخص نظرخواهی میانگین برای مدل سبد کلمات

در جدول ۵-۸ نمونه‌هایی از خروجی مدل به همراه شاخص نظرخواهی میانگین برای آن‌ها ذکر شده است.

تعداد پیشنهادهای مدل	۳۰۰۰			
عبارت توصیفی ورودی	مربوط به سلسله اعصاب			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
عمید	عصبی	امپراتوری	فرهنگ	شاهزاده
شاخص نظرخواهی میانگین	۴.۴	۱.۶	۱	۱.۱
تعداد پیشنهادهای مدل	۵۰۰۰			
عبارت توصیفی ورودی	پاره های یک جسم مرکب را از هم جدا کردن			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
معین	تجزیه	برداشتن	جفت	تقسیم
شاخص نظرخواهی میانگین	۴.۱	۱	۱	۳.۹
تعداد پیشنهادهای مدل	۱۰۰۰۰			
عبارت توصیفی ورودی	راننده هواپیما			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
دهخدا	خلبان	خلبان	سرنشین	هواپیما
شاخص نظرخواهی میانگین	۵	۵	۲.۷	۲.۲
تعداد پیشنهادهای مدل	۲۰۰۰۰			
عبارت توصیفی ورودی	ستیزنده و خصومت کننده			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
دهخدا	متمادی	ستیز	خصومت	دشمنی
شاخص نظرخواهی میانگین	۱	۲.۵	۲.۱	۲.۷

جدول ۵-۸: نمونه‌هایی از خروجی مدل سبد کلمات

۵-۳-۲ مدل حافظه کوتاه-مدت ماندگار

جدول ۵-۹ بهترین تنظیمات مدل حافظه کوتاه-مدت ماندگار را نشان می‌دهد.

اندازه دسته	مدل برداری	بروزرسانی ماتریس جاسازی	داده‌افزایی
۱۶	FT	بله	خیر
تابع خطا	روش بهینه‌سازی	نرخ یادگیری اولیه	پیشگیری از بیش‌برازش
خطای کسینوسی	Adadelta	۱.۰	توقف زودهنگام

جدول ۵-۹: بهترین تنظیمات برای مدل حافظه کوتاه-مدت ماندگار

نتایج ارزیابی مدل روی داده‌های آموزشی و آزمایشی با توجه به سه معیار خطای کسینوسی، دقت در ۱۰ و ۱۰۰ به ترتیب در جدول‌های ۵-۱۰ و ۵-۱۱ آورده شده است.

۳۰۰۰ کلمه		۵۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۵۱		۰.۵۴	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۲۲۵	۰.۵۷۵	۰.۱۵	۰.۴
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۴۲۵	۰.۸۲۵	۰.۳۵	۰.۶۷۵
۱۰۰۰۰ کلمه		۲۰۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۵۵		۰.۵۶	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۱۲۵	۰.۴	۰.۱۲۵	۰.۲۷۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۳	۰.۶۵	۰.۳	۰.۴۵

جدول ۵-۱۰: ارزیابی مدل حافظه کوتاه-مدت ماندگار روی داده‌های آموزشی

۳۰۰۰ کلمه		۵۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۶۱		۰.۶۱	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۱۲۵	۰.۳	۰.۱۲۵	۰.۲
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۲۵	۰.۷	۰.۳۵	۰.۷۲۵
۱۰۰۰۰ کلمه		۲۰۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۶۲		۰.۶۲	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۰۷۵	۰.۲۷۵	۰.۱۵	۰.۲۷۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۳۵	۰.۶۷۵	۰.۴۵	۰.۷۲۵

جدول ۵-۱۱: ارزیابی مدل حافظه کوتاه-مدت ماندگار روی داده‌های آزمایشی

جدول ۵-۱۲ ارزیابی شاخص نظرخواهی میانگین برای مدل حافظه کوتاه-مدت ماندگار را به تفکیک تعداد پیشنهادها نشان می‌دهد.

۳۰۰۰ کلمه				۵۰۰۰ کلمه			
عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۱	۲.۸	۲.۵	۲	۳.۷	۳.۲	۲.۸	۲.۴
معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲.۷	۲	۲.۳	۲.۲	۳.۳	۲.۶	۲.۴	۲.۱
دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲.۱	۱.۹	۱.۸	۱.۶	۲.۴	۲.۳	۱.۸	۱.۹
۱۰۰۰۰ کلمه				۲۰۰۰۰ کلمه			
عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۳	۳.۱	۲.۷	۲.۵	۳.۲	۲.۹	۲.۸	۲.۶
معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۲	۲.۶	۲.۴	۱.۹	۳.۱	۲.۸	۲.۴	۲.۱
دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۱.۹	۲.۵	۲	۲	۱.۹	۲.۲	۱.۸	۱.۷

جدول ۵-۱۲: ارزیابی شاخص نظرخواهی میانگین برای مدل حافظه کوتاه-مدت ماندگار

در جدول ۵-۱۳ چند نمونه از خروجی‌های مدل به همراه شاخص نظرخواهی میانگین برای هر کدام از آن‌ها ذکر شده است.

تعداد پیشنهادهای مدل	۳۰۰۰			
عبارت توصیفی ورودی	انبوهی از مردم			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
عمید	جمعیت	مردم	جمعیت	جامعه
شاخص نظرخواهی میانگین	۴.۸	۲.۷	۴.۸	۳.۴
تعداد پیشنهادهای مدل	۵۰۰۰			
عبارت توصیفی ورودی	جای دور از مردم			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
معین	گوشه	شهر	کوچه	مملکت
شاخص نظرخواهی میانگین	۳.۷	۱.۵	۱.۴	۱.۲
تعداد پیشنهادهای مدل	۱۰۰۰۰			
عبارت توصیفی ورودی	آن رسن که زانوی شتر بدان با میان بندند			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
دهخدا	حجاز	طناب	کوچه	بند
شاخص نظرخواهی میانگین	۱.۱	۳.۸	۲.۷	۲.۵
تعداد پیشنهادهای مدل	۲۰۰۰۰			
عبارت توصیفی ورودی	شکننده و سفید مایل به قرمز به صورت پیرولوزیت در طبیعت فراوان است و در برابر هوا بسیار دیر اکسید می شود			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
معین	منگنز	اکسید	منگنز	نارگیل
شاخص نظرخواهی میانگین	۴.۲	۲.۶	۴.۲	۲

جدول ۵-۱۳: نمونه‌هایی از خروجی مدل حافظه کوتاه-مدت ماندگار

۵-۳-۳ مدل حافظه کوتاه-مدت ماندگار با توجه

جدول ۵-۱۴ بهترین تنظیمات مدل حافظه کوتاه-مدت ماندگار با توجه را نشان می‌دهد.

اندازه دسته	مدل برداری	بروزرسانی ماتریس جاسازی	داده‌افزایی	نوع توجه
۱۶	FT	بله	خیر	جمع‌شونده
تابع خطا	روش بهینه‌سازی	نرخ یادگیری اولیه	پیشگیری از بیش‌برازش	
خطای کسینوسی	Adadelta	۱.۰	توقف زودهنگام	

جدول ۵-۱۴: بهترین تنظیمات برای مدل حافظه کوتاه-مدت ماندگار با توجه

جدول‌های ۵-۱۵ و ۵-۱۶ به ترتیب ارزیابی بهترین مدل با توجه به سه معیار خطای کسینوسی، دقت در ۱۰ و ۱۰۰ روی داده‌های آموزشی و آزمایشی را نشان می‌دهند.

۳۰۰۰ کلمه		۵۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۴۸		۰.۵۲	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۳۲۵	۰.۶۲۵	۰.۲	۰.۴۲۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۴۷۵	۰.۸۲۵	۰.۴۷۵	۰.۷۷۵
۱۰۰۰۰ کلمه		۲۰۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۵۵		۰.۵۷	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۱۷۵	۰.۴۲۵	۰.۱۲۵	۰.۱۷۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۳۷۵	۰.۶۵	۰.۳۵	۰.۴۷۵

جدول ۵-۱۵: ارزیابی مدل حافظه کوتاه-مدت ماندگار با توجه روی داده‌های آموزشی

۳۰۰۰ کلمه		۵۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۶۱		۰.۶۱	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۱۵	۰.۴۲۵	۰.۱۷۵	۰.۲۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۲۷۵	۰.۷	۰.۴	۰.۷۲۵
۱۰۰۰۰ کلمه		۲۰۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۶۲		۰.۶۲	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۱۲۵	۰.۳۲۵	۰.۱۵	۰.۲۲۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۴	۰.۷	۰.۵	۰.۶۲۵

جدول ۵-۱۶: ارزیابی مدل حافظه کوتاه-مدت ماندگار با توجه روی داده‌های آزمایشی

نتایج ارزیابی مدل بر اساس شاخص نظرخواهی میانگین برای بهترین مدل به تفکیک تعداد پیشنهادها در جدول ۵-۱۷ ذکر شده است.

۳۰۰۰ کلمه				۵۰۰۰ کلمه			
عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲.۸	۲.۶	۲.۳	۲.۱	۳.۶	۳	۲.۳	۲.۴
معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲.۷	۲.۳	۲.۲	۱.۹	۳.۷	۲.۷	۲.۱	۱.۹
دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲	۱.۶	۱.۵	۱.۵	۲	۱.۹	۱.۸	۱.۶
۱۰۰۰۰ کلمه				۲۰۰۰۰ کلمه			
عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۲	۳.۱	۲.۸	۲.۷	۳.۱	۲.۶	۲.۳	۲.۱
معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۱	۲.۸	۲.۳	۲.۲	۲.۷	۲.۵	۲.۲	۲
دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲.۱	۲.۲	۲	۱.۹	۱.۹	۲.۲	۲	۱.۷

جدول ۵-۱۷: ارزیابی شاخص نظرخواهی میانگین برای مدل حافظه کوتاه-مدت ماندگار با توجه

جدول ۵-۱۸ نمونه‌هایی از خروجی مدل به همراه شاخص نظرخواهی میانگین برای هر کدام از آن‌ها را نشان می‌دهد.

تعداد پیشنهادهای مدل	۳۰۰۰			
عبارت توصیفی ورودی	تعادل			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
عمید	عدالت	قدرت	سطح	شکل
شاخص نظرخواهی میانگین	۲.۴	۱.۳	۱	۱
تعداد پیشنهادهای مدل	۵۰۰۰			
عبارت توصیفی ورودی	آوازی که از انعکاس صوت به‌وجود می‌آید			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
عمید	صدا	موسیقی	ترانه	آواز
شاخص نظرخواهی میانگین	۴.۴	۳.۴	۲.۸	۲.۹
تعداد پیشنهادهای مدل	۱۰۰۰۰			
عبارت توصیفی ورودی	پول بیگانه			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
معین	ارز	دینار	پولی	خرج
شاخص نظرخواهی میانگین	۴.۷	۴	۲.۳	۲.۱
تعداد پیشنهادهای مدل	۲۰۰۰۰			
عبارت توصیفی ورودی	زیادشده			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
دهخدا	افزودن	قدری	پیچیده	ناگزیر
شاخص نظرخواهی میانگین	۳.۸	۱.۷	۱	۱

جدول ۵-۱۸: نمونه‌هایی از خروجی مدل حافظه کوتاه-مدت ماندگار با توجه

۵-۳-۴ مدل حافظه کوتاه-مدت ماندگار دوسویه با توجه

جدول ۵-۱۹ بهترین تنظیمات مدل حافظه کوتاه-مدت ماندگار دوسویه با توجه را نشان می‌دهد.

اندازه دسته	مدل برداری	بروزرسانی ماتریس جاسازی	داده‌افزایی	نوع توجه
۱۶	FT	بله	خیر	جمع‌شونده
تابع خطا	روش بهینه‌سازی	نرخ یادگیری اولیه	پیشگیری از بیش‌برازش	
خطای کسینوسی	Adadelta	۱.۰	توقف زودهنگام	

جدول ۵-۱۹: بهترین تنظیمات برای مدل حافظه کوتاه-مدت ماندگار با توجه

جدول‌های ۵-۲۰ و ۵-۲۱ ارزیابی بهترین مدل با توجه به سه معیار خطای کسینوسی، دقت در ۱۰ و ۱۰۰ به ترتیب روی داده‌های آموزشی و آزمایشی را نشان می‌دهند.

۳۰۰۰ کلمه		۵۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۴۹		۰.۵۲	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۳۵	۰.۷	۰.۲۲۵	۰.۴۷۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۵	۰.۸۵	۰.۴۷۵	۰.۷۷۵
۱۰۰۰۰ کلمه		۲۰۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۵۴		۰.۵۷	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۱۷۵	۰.۴۲۵	۰.۱۲۵	۰.۲۲۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۴	۰.۶۵	۰.۳۷۵	۰.۵۲۵

جدول ۵-۲۰: ارزیابی مدل حافظه کوتاه-مدت ماندگار دوسویه با توجه روی داده‌های آموزشی

۳۰۰۰ کلمه		۵۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۶۱		۰.۶۱	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۱۵	۰.۳۷۵	۰.۱۷۵	۰.۳
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۲۵	۰.۷۲۵	۰.۳۷۵	۰.۷
۱۰۰۰۰ کلمه		۲۰۰۰۰ کلمه	
مقدار نهایی خطای کسینوسی		مقدار نهایی خطای کسینوسی	
۰.۶۲		۰.۶۲	
Acc_{exact}@۱۰	Acc_{exact}@۱۰۰	Acc_{exact}@۱۰	Acc_{exact}@۱۰۰
۰.۱۲۵	۰.۳۲۵	۰.۱۲۵	۰.۲۲۵
Acc_{syn}@۱۰	Acc_{syn}@۱۰۰	Acc_{syn}@۱۰	Acc_{syn}@۱۰۰
۰.۴۲۵	۰.۷۲۵	۰.۴۵	۰.۶۲۵

جدول ۵-۲۱: ارزیابی مدل حافظه کوتاه-مدت ماندگار دوسویه با توجه روی داده‌های آزمایشی

ارزیابی مدل بر اساس شاخص نظرخواهی میانگین به تفکیک تعداد پیشنهادهای مدل در جدول ۵-۲۲ آورده شده است.

۳۰۰۰ کلمه				۵۰۰۰ کلمه			
عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲.۵	۲.۴	۱.۹	۲.۱	۳.۳	۲.۳	۲.۳	۱.۵
معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۳	۲.۵	۲.۲	۲.۱	۳.۱	۲.۴	۲.۱	۲
دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲.۱	۲	۱.۶	۱.۸	۲.۳	۲.۳	۱.۹	۱.۸
۱۰۰۰۰ کلمه				۲۰۰۰۰ کلمه			
عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	عمید	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۷	۲.۶	۲.۲	۲.۶	۳.۲	۲.۷	۲.۵	۲.۵
معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	معین	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۳.۲	۳.۱	۲.۴	۲	۲.۵	۲.۵	۲	۲.۲
دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم	دهخدا	پیشنهاد اول	پیشنهاد دوم	پیشنهاد سوم
۲.۳	۲.۴	۲.۱	۱.۸	۲.۳	۲.۳	۱.۷	۱.۶

جدول ۵-۲۲: ارزیابی شاخص نظرخواهی میانگین برای مدل حافظه کوتاه-مدت ماندگار دوسویه با توجه

جدول ۵-۲۳ چند نمونه از خروجی‌های مدل و شاخص نظرخواهی میانگین برای هر کدام از آن‌ها را نشان می‌دهد.

تعداد پیشنهادهای مدل	۳۰۰۰			
عبارت توصیفی ورودی	حجت و برهان			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
عمید	دلیل	مسئله	عقیده	استدلال
شاخص نظرخواهی میانگین	۴.۵	۱.۸	۱.۱	۴.۳
تعداد پیشنهادهای مدل	۵۰۰۰			
عبارت توصیفی ورودی	نوشته هایی که با فنون ادبی بستگی داشته و درباره ادبیات باشد			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
معین	ادبی	ادبیات	کتاب	شعر
شاخص نظرخواهی میانگین	۴.۱	۲.۵	۲.۷	۳.۳
تعداد پیشنهادهای مدل	۱۰۰۰۰			
عبارت توصیفی ورودی	مسابقات چند جانبه			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
معین	تورنمنت	تورنمنت	قهرمان	شطرنج
شاخص نظرخواهی میانگین	۴.۸	۴.۸	۳.۵	۲.۴
تعداد پیشنهادهای مدل	۲۰۰۰۰			
عبارت توصیفی ورودی	کنج			
منبع	کلمه صحیح	پیشنهاد اول مدل	پیشنهاد دوم مدل	پیشنهاد سوم مدل
دهخدا	بیخ	گوشه	حیاط	کوچه
شاخص نظرخواهی میانگین	۲.۸	۴.۲	۱	۱

جدول ۵-۲۳: نمونه‌هایی از خروجی مدل حافظه کوتاه-مدت ماندگار دوسویه با توجه

۵-۳-۵ خلاصه یافته‌ها

این قسمت به گزارش مهم‌ترین یافته‌های آزمایش‌ها می‌پردازد. این یافته‌ها در ادامه ذکر می‌شوند.

- استفاده از روش‌های بهینه‌سازی SGD و Adam باعث می‌شود پس از مرحله اول آموزش، خطای مدل روی داده‌های آزمایشی افزایش پیدا کند و در واقع پدیده بیش‌برازش رخ دهد. در حالی که با استفاده از روش Adadelta آموزش پس از مرحله اول ادامه می‌یابد و گاهی به ۷ مرحله نیز می‌رسد. این نتایج مستقل از مقدار نرخ یادگیری صحیح است.
- در تمامی آزمایش‌ها، مدل برداری FT با اختلاف قابل توجه بهتر از مدل‌های دیگر به حل مسئله کمک می‌کند. دلیل این امر می‌تواند پیکره آموزشی بزرگ آن باشد که باعث شده برای اکثریت کلمات موجود در لغت‌نامه‌ها، قابلیت ارائه بردار را داشته باشد.
- با افزایش اندازه دسته، خطای مدل روی داده‌های آزمایشی افزایش کمی می‌یابد.
- بررسی ارزیابی مدل‌ها روی داده‌های آزمایشی با توجه به تعداد پیشنهادها نشان می‌دهد افزایش یا کاهش پیشنهادها، تاثیر قابل توجهی روی مقدار نهایی تابع خطا ندارد. اما بررسی شاخص نظرخواهی میانگین در این موارد تفاوت زیادی را میان عملکرد مدل‌های حاوی حافظه کوتاه-مدت ماندگار و مدل سبد کلمات منعکس می‌کند.
- شاخص نظرخواهی میانگین نشان می‌دهد کیفیت عبارات توصیفی موجود در لغت‌نامه دهخدا، نسبت به دو لغت‌نامه دیگر (معین و عمید) به میزان قابل توجهی پایین‌تر است.
- ارزیابی مدل‌ها روی داده‌های آزمایشی نشان می‌دهد مدل در بسیاری از موارد، به جای کلمه موجود در لیست کلمات صحیح، یکی از کلمات مترادف آن را تحویل داده است.
- داده‌افزایی باعث تسریع وقوع پدیده بیش‌برازش می‌گردد. در تمامی آزمایش‌ها، نمونه‌های آموزشی بدست‌آمده از طریق داده‌افزایی تنها به کاهش خطای مدل روی نمونه‌های آموزشی کمک می‌کنند و تاثیری در خطای مدل روی نمونه‌های آزمایشی ندارند. بررسی نمونه‌های آموزشی ساخته‌شده به این روش نشان می‌دهد این نمونه‌ها اغلب از تغییر عبارات توصیفی بوجود آمده‌اند که مربوط به کلماتی هستند که پیش از این نیز اطلاعات کافی در مورد آن‌ها در داده‌ها موجود بوده است. به نظر می‌رسد کلماتی که توصیف دقیقی از آن‌ها در لغت‌نامه‌ها وجود ندارد، معمولا دارای مترادف نیز نیستند.

فصل ششم

نتایج و کارهای آتی

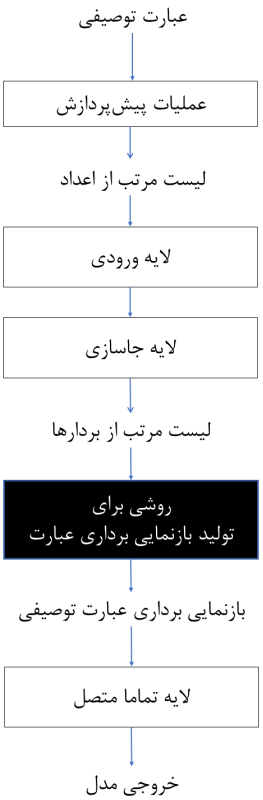
۶-۱ نتایج

در این پایان‌نامه، عملکرد مدل‌های مختلفی بر پایه شبکه‌های عصبی برای تعیین نزدیک‌ترین کلمه به یک مفهوم فارسی مورد بررسی قرار گرفت. فرض اولیه ما این است که کاربر این مفهوم را در قالب یک عبارت (دنباله‌ای از کلمات) توصیف می‌کند. برای آموزش و ارزیابی یک شبکه عصبی که بتواند این عبارت را به یک کلمه نگاشت کند، به داده‌هایی دوتایی از جنس «عبارت توصیفی و کلمه متناظر آن» نیاز داریم. در این پژوهش، این دوتایی‌ها را از طریق لغت‌نامه‌ها، ویکی‌پدیا و شبکه واژگانی فارسی بدست آورده‌ایم.

فرض کنید یک کاربر مفهوم موردنظر خود را به وسیله یک عبارت وصف کرده است. کلمات موجود در این عبارت با نمایشی به شکل یک لیست مرتب از اعداد به عنوان ورودی به هر یک از مدل‌ها داده می‌شوند. معماری کلیه مدل‌ها از یک رویکرد واحد پیروی می‌کند. در هر مدل از یک لایه جاسازی بهره گرفته شده است که وظیفه تبدیل این لیست از اعداد به لیستی از بردارها را بر عهده دارد؛ به طوری که هر بردار متناظر یک کلمه است. لذا ورودی هر مدل پس از عبور از لایه جاسازی، به شکل یک لیست مرتب از بردارها در می‌آید. هر مدل با روشی متفاوت، با پردازش این بردارها، یک بردار جدید تولید می‌کند. از آنجایی که این بردار از روی کلمات موجود در عبارت بدست آمده است، می‌توانیم آن را «بازنمایی برداری عبارت توصیفی» بنامیم. خروجی نهایی هر مدل، یک تصویر از این بازنمایی برداری است. روند تولید خروجی در شکل ۶-۱ به تصویر در آمده است.

برای ارزیابی عملکرد هر مدل روی یک نمونه متشکل از «عبارت توصیفی و کلمه متناظر آن» که از یک لغت‌نامه استخراج شده است، استفاده شد. عبارت موجود در نمونه پس از پیش‌پردازش به شکل یک لیست مرتب از اعداد در آمد و خروجی مدل به ازای آن لیست محاسبه گردید. با مقایسه خروجی و بردارهای متناظر تعداد مشخصی از کلمات زبان فارسی، یک رتبه‌بندی بدست آمد و نزدیک‌ترین کلمات به مفهوم ورودی کاربر تعیین گشت. این کلمات اصطلاحا پیشنهادهای مدل نامیده شدند. برای راحتی در ادامه این بخش از کلمه موجود در هر نمونه - از آنجایی که از یک لغت‌نامه استخراج شده است - با عنوان «کلمه پیشنهادی لغت‌نامه» یاد می‌کنیم.

در مدل «سبد کلمات»، بازنمایی برداری عبارت توصیفی با محاسبه میانگین بردارهای کلمات موجود در عبارت بدست می‌آید. لذا با اتخاذ این رویکرد، ترتیب قرارگیری کلمات در بازنمایی برداری عبارت لحاظ نمی‌شود. نتایج ارزیابی این مدل بر اساس شاخص نظرخواهی میانگین نشان می‌دهد که در اغلب موارد، مدل کلمه‌ای غیر مرتبط را به ازای دریافت عبارت توصیفی تحویل می‌دهد و لذا ناتوان از تامین



شکل ۶-۱: رویکرد کلی مدل‌های پیشنهادی

خواسته کاربر است.

مدل «حافظه کوتاه-مدت ماندگار»، از همین نوع شبکه عصبی برای محاسبه بازنمایی برداری عبارت استفاده می‌کند. این مدل بر خلاف مدل سبد کلمات، بردارها را به ترتیب پردازش می‌کند و به همین دلیل، ترتیب رخداد کلمات نیز در بازنمایی برداری عبارت تاثیرگذار است. این مدل از نظر میزان خطای کسینوسی روی داده‌های آزمایشی تفاوتی کمتر از ۰.۰۳ با مدل سبد کلمات دارد؛ اما معیارهای ارزیابی دقت در k و شاخص نظرخواهی میانگین، برتری آن نسبت به مدل سبد کلمات را منعکس می‌کنند. همچنین شاخص نظرخواهی میانگین نشان می‌دهد پیشنهادهای این مدل، از نظر میزان کیفیت قابل مقایسه با کلمات پیشنهادی لغت‌نامه هستند.

در مدل «حافظه کوتاه-مدت ماندگار با توجه»، اهمیت متفاوتی برای هر کلمه در نظر گرفته می‌شود؛ به طوری که بعضی از کلمات نقش بیشتری در تولید بازنمایی برداری عبارت دارند. یک لایه توجه از نوع جمع‌شونده این قابلیت را برای مدل فراهم می‌کند. با این وجود، این مدل از نظر میزان خطای کسینوسی روی داده‌های آموزشی بسیار مشابه مدل «حافظه کوتاه-مدت ماندگار» عمل می‌کند. شاخص نظرخواهی میانگین نیز نشان می‌دهد که پیشنهادهای این مدل از کیفیت مناسبی برخوردار هستند. در

موارد معدودی، این پیشنهادها با توجه شاخص نظرخواهی میانگین، از کیفیت بهتری نسبت به کلمات پیشنهادی لغت‌نامه دهخدا برخوردار هستند.

مدل «حافظه کوتاه-مدت ماندگار دوسویه با توجه»، عبارت را یک بار از ابتدا به انتها و بار دیگر از انتها به ابتدا پردازش می‌کند. از لحاظ تئوری، این پردازش دوسویه می‌تواند از توجه کمتر مدل به کلمات انتهایی عبارت به دلیل وقوع پدیده کاهش گرادیان جلوگیری کند؛ زیرا هر کلمه که در انتهای عبارت اصلی ظاهر شده باشد، با پردازش از انتها به ابتدا، در ابتدای عبارت جدید قرار می‌گیرد. بر خلاف تصور اولیه، این مدل بر اساس معیار ارزیابی دقت در k حداکثر ۰.۰۵ بهتر یا بدتر از مدل «حافظه کوتاه-مدت ماندگار با توجه» عمل می‌کند و از نظر میزان خطای کسینوسی روی داده‌های آزمایشی کمتر از ۰.۰۱ با آن تفاوت دارد. شاخص نظرخواهی میانگین نشان می‌دهد پیشنهادهای مدل «حافظه کوتاه-مدت ماندگار دوسویه با توجه» نیز کیفیتی قابل مقایسه با کلمات پیشنهادی لغت‌نامه‌ها دارند و تنها در موارد معدودی از آن‌ها بهتر هستند که این موارد در مورد کلمات پیشنهادی لغت‌نامه دهخدا اتفاق می‌افتد. لذا در مجموع، با افزودن قابلیت پردازش دوسویه به مدل «حافظه کوتاه-مدت ماندگار با توجه»، پیشرفت واضحی در عملکرد آن مدل مشاهده نمی‌شود.

در مجموع ارزیابی مدل‌های مختلف پیشنهادی نشان می‌دهد که مدل‌های «سبد کلمات» و «حافظه کوتاه-مدت ماندگار» عملکرد به مراتب ضعیف‌تری نسبت به دو مدل دیگر دارند. این در حالی است که دو مدل «حافظه کوتاه-مدت ماندگار با توجه» و «حافظه کوتاه-مدت ماندگار دوسویه با توجه» تفاوت محسوسی از نظر کیفیت پیشنهادها ندارند. در نتیجه با در نظر داشتن پیچیدگی بیشتر مدلی که از حافظه کوتاه-مدت ماندگار دوسویه استفاده می‌کند، استفاده از آن برای حل مسئله مورد بررسی در این پایان‌نامه به صرفه نیست. لذا ما در این پژوهش مدل «حافظه کوتاه-مدت ماندگار با توجه» را برای تعیین نزدیک‌ترین کلمه به یک مفهوم فارسی پیشنهاد می‌کنیم.

۶-۲ کارهای آتی

در بخش نتایج، نتایج بررسی مدل‌های پیشنهادی ما برای حل مسئله بیان گردید. در این بخش، رویکردهای دیگری برای حل مسئله و روش‌هایی برای بهبود مدل‌های پیشنهادی مطرح می‌گردد که می‌تواند به عنوان پژوهش‌های آتی در نظر گرفته شود.

۶-۲-۱ استفاده از بردارهای متفاوت برای معانی مختلف کلمات

معانی مختلف یک کلمه، حاوی مفاهیم متفاوتی هستند. به طور مثال، کلمه «سلام» غالبا به عنوان «واژه‌ای در آغاز گفتگو» به کار برده می‌شود؛ اما می‌تواند به عنوان «احترام نظامی» نیز در نظر گرفته شود. بررسی لغت‌نامه عمید نشان می‌دهد این کلمه در گذشته‌های دور به معنی «رهایی از عیب و آفت» نیز کاربرد داشته است. در مدل‌هایی که تا کنون برای بازنمایی برداری کلمات فارسی ارائه شده‌اند، این تفاوت‌ها لحاظ نشده است. از طرفی، این مدل‌ها غالبا با استفاده از متون امروزی فارسی آموزش داده شده‌اند؛ به همین دلیل قابلیت کمتری برای تمایز میان معانی قدیمی و جدید دارند. با توجه به این مشکل، اگر بردارهایی مخصوص هر معنی کلمات فارسی ارائه شود، می‌توان از آن بردارها برای تشکیل ماتریس‌های جاسازی و مقایسه در این پژوهش استفاده نمود. با توجه به قابلیت بردارهای مذکور در تمایز میان معانی، انتظار می‌رود به کارگیری این بردارها باعث بهبود کارایی مدل‌ها گردد.

۶-۲-۲ پیش‌بینی جزء کلام و حوزه معنایی کلمه متناظر عبارت توصیفی کاربر

اگر سیستمی طراحی شود که با دریافت یک عبارت توصیفی، جزء کلام یا حوزه معنایی کلمه متناظر آن را مشخص نماید، می‌توان با ترکیب خروجی این مدل با خروجی مدل‌های پیشنهادی در این پژوهش، مسئله را به طور دقیق‌تر حل نمود و پیشنهادهای مرتبط‌تری را به کاربر ارائه کرد. مثلا اگر کاربر عبارت «کسی که در مزرعه کار می‌کند» را در نظر داشته باشد، این سیستم می‌تواند اعلام کند که کلمه مورد نظر (کشاورز) یک «اسم» است. همچنین اگر عبارت «جایی که نمایندگان قشرهای مختلف برای تصویب قوانین در آن جمع می‌شوند» مد نظر کاربر باشد، سیستم می‌تواند حوزه «سیاسی» را اعلام نماید. با داشتن این اطلاعات اضافه می‌توان جستجوی بهتری را در میان کلمات فارسی برای یافتن نزدیک‌ترین کلمه به عبارت موردنظر کاربر انجام داد. مثلا می‌توان به کلماتی که دارای برچسب «سیاسی» هستند،

امتیاز بیشتری نسبت داد. متاسفانه حجم داده‌های برچسب‌دار موجود در لغت‌نامه‌ها بسیار کمتر از کل داده‌ها است و توزیع نامتعادل برچسب‌های آن‌ها نیز برای آموزش چنین سیستم‌هایی مشکل‌ساز است. به طور مثال، تنها ۳۶۹۶ عبارات دارای برچسب «اصطلاح» هستند و ۵۳۶۷ عبارت دارای برچسب «قید» هستند؛ در حالی که ۱۱۷۲۳۱ عبارت برچسب «اسم» دارند. این توزیع نامتعادل می‌تواند باعث سوگیری [۱] مدل به سمت برچسب‌های با تکرار بیشتر شود. در جریان انجام آزمایش‌های این پژوهش، با استفاده از مدل‌های مختلفی از جمله شبکه‌های عصبی پیچشی، سعی در آموزش سیستمی برای پیش‌بینی برچسب‌های جزء کلام و حوزه معنایی برای کلمه متناظر یک عبارت توصیفی نمودیم. به دلیل نتایج بسیار ضعیف این مدل‌ها، از ذکر جزئیات این مدل‌ها خودداری می‌کنیم.

۶-۲-۳ استفاده از بازنمایی‌های رمزگذار دوسویه بر پایه تبدیل‌کننده

در سال‌های اخیر، مدل‌هایی تحت عنوان «بازنمایی‌های رمزگذار دوسویه بر پایه تبدیل‌کننده» ارائه شده‌اند که با از استفاده از شبکه‌های عصبی عمیق، متون را بازنمایی می‌کنند. نتایج آزمایش‌ها نشان می‌دهد اگر به معماری این مدل‌ها یک لایه خروجی بر اساس شرایط مسئله اضافه شود، قابلیت استفاده برای حل مسائل مختلف در حوزه پردازش زبان طبیعی را دارند [۱۱]. به تازگی یکی از این مدل‌ها با نام ParsBERT توسعه داده شده است که مخصوص زبان فارسی می‌باشد [۲۷]. با تنظیم دقیق [۲] وزن‌های این مدل و مدل‌های مشابه پس از افزودن لایه خروجی، این امکان وجود دارد که به نتایج بهتری در زمینه تعیین نزدیک‌ترین کلمه به یک مفهوم فارسی دست یابیم.

ذکر این نکته حائز اهمیت است که پیچیدگی این مدل‌ها، بسیار فراتر از مدل‌هایی است که در این پایان‌نامه پیشنهاد کرده‌ایم. کلیه مدل‌های ارائه‌شده توسط ما دارای کمتر از ۳۲ میلیون وزن قابل یادگیری هستند؛ در حالی که به طور مثال مدل ParsBERT حاوی بالغ بر ۱۶۲ میلیون وزن قابل یادگیری می‌باشد. لذا آموزش این گونه مدل‌ها نیاز به صرف زمان بیشتر و تجهیزات سخت‌افزاری پیشرفته‌تری دارد.

[۱] Bias
[۲] Fine-tuning

منابع و مراجع

[1] AleAhmad, Abolfazl, Amiri, Hadi, Darrudi, Ehsan, Rahgozar, Masoud, and Oroumchian, Farhad. Hamshahri: A standard persian text collection. *Knowledge-Based Systems*, 22(5):382–387, 2009.

[2] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Bengio, Yoshua, Ducharme, Réjean, Vincent, Pascal, and Jauvin, Christian. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[4] Bijankhan, Mahmood. The role of the corpus in writing a grammar: An introduction to a software. *Iranian Journal of Linguistics*, 19(2):48–67, 2004.

[5] Bilac, Slaven, Watanabe, Wataru, Hashimoto, Taiichi, Tokunaga, Takenobu, and Tanaka, Hozumi. Dictionary search based on the target word description. in *Proc. of the Tenth Annual Meeting of The Association for Natural Language Processing (NLP2004)*, pp. 556–559, 2004.

[6] Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, and Mikolov, Tomas. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[7] Brown, Roger and McNeill, David. The “tip of the tongue” phenomenon. *Journal of verbal learning and verbal behavior*, 5(4):325–337, 1966.

[8] Cicchetti, Domenic V and Allison, Truett. A new procedure for assessing reliability of scoring eeg sleep recordings. *American Journal of EEG Technology*, 11(3):101–110, 1971.

[9] Cohen, Jacob. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[10] Cohen, Jacob. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[11] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Dutoit, Dominique and Nugues, Pierre. A lexical database and an algorithm to find words from definitions. in *ECAI*, pp. 450–454, 2002.

[13] El-Kahlout, Ilknur Durgar and Oflazer, Kemal. Use of wordnet for retrieving words from their meanings. in *Proceedings of the global Wordnet conference (GWC2004)*, pp. 118–123, 2004.

[14] Engelbrecht, Andries P. *Computational intelligence: an introduction*. John Wiley & Sons, 2007.

[15] Ferret, Olivier. Using collocations for topic segmentation and link detection. in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[16] Ferret, Olivier and Zock, Michael. Enhancing electronic dictionaries with an index based on associations. in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 281–288, 2006.

[17] Freeman, Linton C. A set of measures of centrality based on betweenness. *Sociometry*, pp. 35–41, 1977.

[18] Grave, Edouard, Bojanowski, Piotr, Gupta, Prakhar, Joulin, Armand, and Mikolov, Tomas. Learning word vectors for 157 languages. in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[19] Hadifar, Amir and Momtazi, Saeedeh. The impact of corpus domain on word representation: a study on persian word embeddings. *Language Resources and Evaluation*, 52(4):997–1019, 2018.

[20] Hedderich, Michael A, Yates, Andrew, Klakow, Dietrich, and de Melo, Gerard. Using multi-sense vector embeddings for reverse dictionaries. *arXiv preprint arXiv:1904.01451*, 2019.

[21] Hill, Felix, Cho, KyungHyun, Korhonen, Anna, and Bengio, Yoshua. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30, 2016.

[22] Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[23] Kim, Yoon. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[24] Krogh, Anders and Hertz, John A. A simple weight decay can improve generalization. in *Advances in neural information processing systems*, pp. 950–957, 1992.

[25] Landis, J Richard and Koch, Gary G. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.

[26] Mandal, S and Prabaharan, N. Ocean wave forecasting using recurrent neural networks. *Ocean engineering*, 33(10):1401–1410, 2006.

[27] Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani Mohammad Manthouri. Parsbert: Transformer-based model for persian language understanding. *ArXiv*, abs/2005.12515, 2020.

[28] Méndez, Oscar, Calvo, Hiram, and Moreno-Armendáriz, Marco A. A reverse dictionary based on semantic analysis using wordnet. in *Mexican International Conference on Artificial Intelligence*, pp. 275–285. Springer, 2013.

[29] Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[30] Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff. Distributed representations of words and phrases and their compositionality. in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[31] Mikolov, Tomas and Zweig, Geoffrey. Context dependent recurrent neural network language model. in *2012 IEEE Spoken Language Technology Workshop (SLT)*, pp. 234–239. IEEE, 2012.

[32] Miller, George A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[33] Morinaga, Yuya and Yamaguchi, Kazunori. Improvement of reverse dictionary by tuning word vectors and category inference. in *International Conference on Information and Software Technologies*, pp. 533–545. Springer, 2018.

[34] Page, Lawrence, Brin, Sergey, Motwani, Rajeev, and Winograd, Terry. The pagerank citation ranking: Bringing order to the web. tech. rep., Stanford InfoLab, 1999.

[35] Pelletier, Francis Jeffry. The principle of semantic compositionality. *Topoi*, 13(1):11–24, 1994.

[36] Pilehvar, Mohammad Taher. On the importance of distinguishing word meaning representations: A case study on reverse dictionary mapping. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2151–2156, 2019.

[37] Pilehvar, Mohammad Taher and Collier, Nigel. De-conflated semantic representations. *arXiv preprint arXiv:1608.01961*, 2016.

[38] Prechelt, Lutz. Early stopping-but when? in *Neural Networks: Tricks of the trade*, pp. 55–69. Springer, 1998.

[39] Rather, Akhter Mohiuddin, Agarwal, Arun, and Sastry, VN. Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications*, 42(6):3234–3241, 2015.

[40] Reyes-Magaña, Jorge, Bel-Enguix, Gemma, Sierra, Gerardo, and Gómez-Adorno, Helena. Designing an electronic reverse dictionary based on two word association norms of english language. *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography*, p. 142, 2019.

[41] Salehinejad, Hojjat, Sankar, Sharan, Barfett, Joseph, Colak, Errol, and Valaee, Shahrokh. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

[42] Schuster, Mike and Paliwal, Kuldip K. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[43] Seraji, Mojgan. *Morphosyntactic corpora and tools for Persian*. Ph.D. thesis, Acta Universitatis Upsaliensis, 2015.

[44] Shamsfard, Mehrnoush, Hesabi, Akbar, Fadaei, Hakimeh, Mansoory, Niloofar, Famian, Ali, Bagherbeigi, Somayeh, Fekri, Elham, Monshizadeh, Maliheh, and Assi, S Mostafa. Semi automatic development of farsnet; the persian wordnet. vol. 29, 2010.

[45] Shaw, Ryan, Datta, Anindya, VanderMeer, Debra, and Dutta, Kaushik. Building a scalable database-driven reverse dictionary. *IEEE Transactions on Knowledge and Data Engineering*, 25(3):528–540, 2011.

[46] Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[47] Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural networks. in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[48] The Pennsylvania State University. Cohenś kappa statistic for measuring agreement, 2018. https://online.stat.psu.edu/stat509/node/162/.

[49] Thorat, Sushrut and Choudhari, Varad. Implementing a reverse dictionary, based on word definitions, using a node-graph architecture. *arXiv preprint arXiv:1606.00025*, 2016.

[50] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia. Attention is all you need. in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[51] Zheng, Lei, Qi, Fanchao, Liu, Zhiyuan, Wang, Yasheng, Liu, Qun, and Sun, Maosong. Multi-channel reverse dictionary model. in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 312–319, 2020.

[52] Zock, Michael and Bilac, Slaven. Word lookup on the basis of associations: from an idea to a roadmap. in *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pp. 29–35, 2004.

[53] Zock, Michael and Schwab, Didier. Lexical access based on underspecified input. in *COLING 2008: Proceedings of the Workshop on cognitive Aspects of the Lexicon (COGALEX 2008)*, pp. 9–17, 2008.

[۵۴] خداپرستی، فرج‌الله. فرهنگ جامع واژگان مترادف و متضاد زبان فارسی. دانشنامه فارس، ۱۳۷۶.

[۵۵] طوسی، اسدی. لغت فُرس. پیرایش عباس اقبال. با سرمایه عبدالرحیم خلخالی. تهران: چاپخانه مجلس، ۱۳۱۹.

[۵۶] کشانی، خسرو. فرهنگ فارسی زانسو. مرکز نشر دانشگاهی، ۱۳۷۲.

پیوست

در این پایان‌نامه، از واژه‌نامه تخصصی هوش مصنوعی [۳] که در اسفند ۱۳۹۵ در آزمایشگاه پردازش داده‌ها و علائم دانشکده علوم و فنون نوین دانشگاه تهران انتشار یافته است، به عنوان منبعی برای انتخاب کلمات معادل برای اصطلاحات تخصصی انگلیسی بهره برده‌ایم.

کلیه داده‌ها، کدهای برنامه‌نویسی و مدل‌های بدست‌آمده در این پژوهش، پس از انتشار پایان‌نامه در تارنمای Github [۴] قابل دسترسی خواهد بود.

[۳] http://dsp.ut.ac.ir/en/wp-content/uploads/2017/05/DSP-Dictionary.ver1_.pdf

[۴] https://github.com/arm-on

واژه‌نامه‌ی فارسی به انگلیسی

آ

آستانه . . . . . Threshold

اصل ترکیب معنایی . . . . . The principle of semantic compositionality

امتیاز توافق . . . . . Agreement score

انفجار گرادیان . . . . . Exploding gradient

ایست‌واژه . . . . . Stopword

ب

بازنمایی برداری کلمه . . . . . Word vector representation

بازنمایی‌های رمزگذار دوسویه بر پایه تبدیل‌کننده . . . . . Bidirectional encoder representations from transformers (BERT)

بردار بایاس . . . . . Bias vector

بیش‌برازش . . . . . Overfit

پ

پایگاه داده ذخیره‌شده . . . . Database dump

پرسش . . . . . Query

پرسش و پاسخ . . . . . Question answering

پیکره . . . . . Corpus

پیوند درونی . . . . . Interlink

ت

تابع فعال‌سازی . . . . . Activation function

تابع هزینه . . . . . Cost function

تبدیل‌کننده . . . . . Transformer

ترادف . . . . . Synonymy

تصویر . . . . . Projection

تعمیم‌پذیری . . . . . Generalization

تقسیم‌بندی موضوع . . . Topic segmentation

تماما متصل . . . . . Fully connected

تنزل وزن . . . . . Weight decay

تنظیم دقیق . . . . . Fine-tuning

توافق مورد انتظار . . . . Expected agreement

توافق نسبی مشاهده‌شده . . . . . Observed proportional agreement

توقف زودهنگام . . . . . Early stopping

ج

جاسازی . . . . . Embedding

جزء کلام . . . . . Part of speech

جمع‌شونده . . . . . Additive

ح

حافظه کوتاه-مدت ماندگار . Long short-term memory

حافظه کوتاه-مدت ماندگار دوسویه Bidirectional long short-term memory

حالت مخفی . . . . . Hidden State

حذف تصادفی . . . . . Dropout

خ

خزنده وب . . . . . Web crawler

د

داده‌افزایی . . . . . Data Augmentation

دسته . . . . . Batch

ز

زیرشمولی . . . . . Hyponymy

س

سبد n-تایی های کاراکتری Bag of character n-grams

سرواژه . . . . . Lemma

سوگیری . . . . . Bias

ش

شاخص نظرخواهی میانگین . Mean Opinion Score

شبکه عصبی بازگشتی . . . Recurrent neural network

شبکه عصبی پیچشی . Convolutional neural network

شبکه واژگانی . . . . . WordNet

ض

ضرب نقطه‌ای . . . . . Dot-product

ف

فراشمولی . . . . . Hypernymy

ک

کدگذاری . . . . . Encoding

گ

گیت خروجی . . . . . Output gate

گیت فراموشی . . . . . Forget gate

گیت ورودی . . . . . Input gate

ل

لایه . . . . . Layer

لغت‌نامه کامل . . . . . Integral Dictionary

م

ماتریس وزن . . . . . Weight matrix

مترادف . . . . . Synonym

محو گرادیان . . . . . Vanishing gradient

مدل زبانی . . . . . Language model

مرحله . . . . . Epoch

معنی کلمه . . . . . Word sense

مقیاس‌پذیر . . . . . Scaled

مکانیزم توجه . . . . . Attention mechanism

موجودیت . . . . . Entity

ن

نرمال‌سازی . . . . . Normalization

و

واحد . . . . . Token

واحدسازی . . . . . Tokenization

وب سرویس . . . . . Web service

ویژگی معنایی . . . . . Semantic feature

ه

هنجارهای انجمنی . . . . Association Norms

واژه‌نامه‌ی انگلیسی به فارسی

A

Activation function . . . . . تابع فعال‌سازی

Additive . . . . . جمع‌شونده

Agreement score . . . . . امتیاز توافق

Association Norms . . . . هنجارهای انجمنی

Attention mechanism . . . . . مکانیزم توجه

B

Bag of character n-grams سبد n-تایی های کاراکتری

Batch . . . . . دسته

Bias . . . . . سوگیری

Bidirectional encoder representations from transformers (BERT) بازنمایی‌های رمزگذار دوسویه بر پایه تبدیل‌کننده . . . . . -

Bias vector . . . . . بردار بایاس

حافظه کوتاه-مدت ماندگار دوسویه Bidirectional long short-term memory

C

Convolutional neural network شبکه عصبی پیچشی .

Corpus . . . . . پیکره

Cost function . . . . . تابع هزینه

D

Data Augmentation . . . . . داده‌افزایی

Database dump . . . . پایگاه داده ذخیره‌شده

Dot-product . . . . . ضرب نقطه‌ای

Dropout . . . . . حذف تصادفی

E

Early stopping . . . . . توقف زودهنگام

Embedding . . . . . جاسازی

Encoding . . . . . کدگذاری

Entity . . . . . موجودیت

Epoch . . . . . مرحله

Expected agreement . . . . توافق مورد انتظار

Exploding gradient . . . . . انفجار گرادیان

F

Fine-tuning . . . . . تنظیم دقیق

Forget gate . . . . . گیت فراموشی

Fully connected . . . . . تماما متصل

G

Generalization . . . . . تعمیم‌پذیری

H

Hidden State . . . . . حالت مخفی

Hypernymy . . . . . فراشمولی

Hyponymy . . . . . زیرشمولی

I

Integral Dictionary . . . . . لغت‌نامه کامل

Interlink . . . . . پیوند درونی

L

Language model . . . . . مدل زبانی

Layer . . . . . لایه

Lemma . . . . . سرواژه

Long short-term memory . حافظه کوتاه-مدت ماندگار

M

Mean Opinion Score . شاخص نظرخواهی میانگین

N

Normalization . . . . . نرمال‌سازی

O

Observed proportional agreement . . . . . توافق نسبی مشاهده‌شده

Output gate . . . . . گیت خروجی

Overfit . . . . . بیش‌برازش

P

Part of speech . . . . . جزء کلام

Projection . . . . . تصویر

Q

Query . . . . . پرسش

Question answering . . . . . پرسش و پاسخ

R

تقسیم‌بندی موضوع . . Topic Segmentation

تبدیل‌کننده . . . . . Transformer

V

محو گرادیان . . . . . Vanishing gradient

W

خزنده وب . . . . . Web crawler

وب سرویس . . . . . Web service

تنزل وزن . . . . . Weight decay

ماتریس وزن . . . . . Weight matrix

شبکه واژگانی . . . . . WordNet

معنی کلمه . . . . . Word sense

بازنمایی برداری کلمه . . . . . Word vector representation

شبکه عصبی بازگشتی . . . Recurrent neural network

S

مقیاس‌پذیر . . . . . Scaled

ویژگی معنایی . . . . . Semantic feature

ایست‌واژه . . . . . Stopword

مترادف . . . . . Synonym

ترادف . . . . . Synonymy

T

اصل ترکیب معنایی . . . . . The principle of semantic compositionality

آستانه . . . . . Threshold

واحد . . . . . Token

واحدسازی . . . . . Tokenization

Abstract

Accessing the appropriate words to convey the concepts, i.e., lexical access is essential for effective communication among people. Considering the computer simulation of the lexical access, we could think about a system that allows the user to find the words which could relate to a specific concept. Such a system is called a reverse dictionary. In this thesis, we presented different models for implementing an intelligent reverse dictionary for Persian language using artificial neural networks. These models should be able to map a descriptive phrase to a word corresponding to it. Therefore, we used *(phrase, word)* tuples in order to train them. These tuples were extracted from three Persian language resources: Dictionaries (Amid, Dehkhoda and Moin), Wikipedia, and Wordnet. Each model should be able to map all of the phrases extracted from different sources describing a word to itself. Also, since Amid and Moin dictionaries were both published more than thirty years after the Dehkhoda dictionary, different word choice and sentence structure is evident when comparing these resources. The model is also expected to handle these differences. A proportion of the tuples extracted from the dictionaries were used to evaluate the functionality of the proposed system. The lexical entry of each word in these dictionaries contains the phrases which describe it. So, the word corresponding with each lexical entry could be considered as the nearest word to the descriptive phrases, based on the judgement of the dictionary's developer. Our experiments revealed that with choosing the right word vector representation and using a Long Short-Term Memory (LSTM) along with attention and dense layers, we could design a model well capable of implementing a reverse dictionary. More precisely, the output words generated by the model are comparable/analogous with (or in same cases better than) the word in the original dictionary in terms of the ability to convey the concepts within the input phrase.

Key Words:

Reverse dictionary, Concept search, Lexical access, Natural language processing, Artificial neural networks

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Mathematics and Computer Science

M. Sc. Thesis

Determining the closest neighboring word to a concept using a Persian reverse dictionary

By

Arman Malekzadeh Lashkaryani

Supervisors

Dr. Ali Mohades Khorasani and Dr. Amin Gheibi

Advisor

Dr. Mohammad Bahrani

August 2020