
بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

رساله دکتری مهندسی هوش مصنوعی

بهبود ارزیابی شباهت متن با استفاده از روش‌های آماری و معیارهای مبتنی بر شبکه واژگان

نگارنده: رضا جوادزاده

استاد راهنما

دکتر مرتضی زاهدی

استاد مشاور

دکتر مرضیه رحیمی

تیر ۱۴۰۰

شماره:
تاریخ:
ویرایش:

باسمه تعالی

مدیریت تحصیلات تکمیلی

فرم شماره 11: صورت جلسه نهایی دفاع از رساله دکتری (Ph.D)

بدینوسیله گواهی می شود آقای/خانم رضا جوادزاده دانشجوی دکتری رشته **مهندسی کامپیوتر** به شماره دانشجویی **۹۶۰۰۷۶۵** ورودی مهر ماه سال ۱۳۹۶ در تاریخ ۱۴۰۰/۴/۰۹ از رساله نظری■ / عملی□ خود با عنوان: **بهبود ارزیابی شباهت متن با استفاده از روش‌های آماری و معیارهای مبتنی بر شبکه واژگان** دفاع و با اخذ نمره ۱۷ به درجه : خوب نائل گردید.

جدول تعیین درجه نمره رساله برای ورودیهای 94 و ماقبل

الف) درجه عالی: نمره 20-19 □	ب) درجه خیلی خوب: نمره 18/99 – 17□
ج) درجه خوب: نمره16/99– 15□	د) مردود: کمتر از 15 □

جدول تعیین درجه نمره رساله برای ورودیهای 95 و مابعد

الف) درجه عالی: نمره 20-19 □	ب) درجه خیلی خوب: نمره 18/99 – 18□
ج) درجه خوب: نمره17/99– 16☑	د) مردود: کمتر از 16 □

ردیف	هیئت داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
1	دکتر مرتضی زاهدی	استاد/ اساتید راهنما	استادیار	
2	دکتر مرضیه رحیمی	مشاور/ مشاورین	استادیار	
3	دکتر بهروز مینایی	استاد مدعو خارجی	دانشیار	
4	دکتر حسین مروی	استاد مدعو داخلی	دانشیار	
5	دکتر هدی مشایخی	استاد مدعو داخلی	استادیار	
	دکتر علیرضا تجری	سرپرست (نماینده) تحصیلات تکمیلی دانشکده	استادیار	

مدیر محترم تحصیلات تکمیلی دانشگاه:

ضمن تأیید مراتب فوق مقرر فرمائید اقدامات لازم در خصوص انجام مراحل دانش آموختگی آقای **رضا جوادزاده** بعمل آید.

نام و نام خانوادگی رئیس دانشکده: منصور رستم

تاریخ و امضاء و مهر دانشکده: ۱۴۰۰/۶/۸

وزارت علوم، تحقیقات و فناوری
دانشگاه فناوری اطلاعات و ارتباطات

تقدیم به آن‌ها که آموختند مرا تا بیاموزم

استادان گرامی

دکتر زاهدی

دکتر رحیمی

دکتر مشایخی

دکتر فاتح

تشکر و قدردانی

از جناب آقای دکتر زاهدی

که آموزش‌های ایشان ارزش فراگیری در این مقطع را دوچندان نمود

تعهد نامه

اینجانب رضا جوادزاده دانشجوی دوره دکتری رشته مهندسی کامپیوتر- هوش مصنوعی دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده پایان‌نامه بهبود ارزیابی شباهت متن با استفاده از روش‌های آماری و معیارهای مبتنی بر شبکه واژگان تحت راهنمائی دکتر مرتضی زاهدی متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام «دانشگاه صنعتی شاهرود» و یا «Shahrood University of Technology » به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

چکیده

هرروزه، حجم بالایی از اطلاعات متنی در فضای اینترنت منتشر می‌شود. علاوه بر این، کتاب‌های بسیاری از فرهنگ‌های مختلف، به صورت دیجیتالی عرضه شده است و در کتابخانه‌های دیجیتال نگهداری می‌شود. مقدار وسیعی از این داده‌ها به صورت زبان طبیعی تولید می‌شوند. این باعث می‌شود که اهمیت استخراج اطلاعات مهم از حجم بالای داده‌ها پررنگ‌تر شود. محاسبه شباهت زوج‌جمله و دقت این ارزیابی، مسئله مهمی با کاربردهای فراوان است. جستجوی معنایی، خلاصه‌سازی استخراجی، ارزیابی احساسات، رده‌بندی اسناد و تشخیص سرقت ادبی همه می‌توانند به عنوان مسئله شباهت زوج‌جمله مدل شوند. ارزیابی شباهت به محاسبه شباهت عبارت‌های بین دو جمله می‌پردازد. موضوع پژوهش پیش رو، ارزیابی شباهت بین دو جمله است در نتیجه به صورت پیش‌فرض، طول بین دو جمله محدود به کمتر از دو خط در نظر گرفته می‌شود. در مسئله شباهت زوج‌جمله، دو جمله به مدل طراحی شده ارسال می‌شود و سیستم باید عددی بین صفر و یک تولید کند (صفر یعنی دو جمله اصلاً شبیه نیستند و یک یعنی دو جمله به صورت تام شباهت معنایی دارند). هدف ما طراحی روشی مبتنی بر رگرسیون است به صورتی که بیشترین همبستگی را با نمره‌های تخصیص داده‌شده توسط داوران انسانی داشته باشد.

در روش پیشنهادی تبدیل موجک، ابتدا توصیف‌گر روی داده‌های آموزش بهینه‌سازی می‌شود، سپس روی توصیف جمله تبدیل موجک گرفته شده است؛ در ادامه، «شباهت کاسینوسی کانال‌های متناظر بین دو جمله» به عنوان «ویژگی‌های پیش‌بینی ارتباط زوج‌جمله» در نظر گرفته می‌شود. نتایج نشان‌دهنده بهبود عملکرد نسبت به روش RoBERTa-base می‌باشد و همچنین مقدار همبستگی نسبت به تمامی روش‌های base که مدل‌های پایه هستند، بهبود یافته است. روش SimCSE از رویکرد ما عملکرد بهتری دارد اما در مقایسه، مدل توصیف‌گر مورد استفاده ما کوچکتر از مدل روش SimCSE است و در نتیجه زمان آموزش و آزمون در آن کمتر می‌باشد.

کلمات کلیدی: شباهت متن، هم‌وقوعی واژه‌ها، پیچیدگی متن، تطابق رشته‌ها، شباهت واژه‌ها، رگرسیون.

فهرست مقاله‌های مستخرج از رساله

۱. در این پژوهش روش پیشنهادی توضیح داده شده در بخش ۳-۱۳-۵ آورده شده است. این روش کاملاً مبتنی بر اطلاعات شبکه واژگان است و دو روش متفاوت (۱) ماتریس شباهت زوج‌واژگان جمله‌ها و (۲) تشکیل بردار ویژگی برای هرجمله، محاسبه شده‌اند که با همدیگر تلفیق می‌شوند.

Reza Javadzadeh, Morteza Zahedi, Marziea Rahimi. ”Sentence similarity using weighted path and similarity matrices”. Turkish Journal of Electrical Engineering and Computer Sciences: Accepted, 2020

۲. در این پژوهش یکی از روش‌های اخیر در زمینه «شباهت مبتنی بر شبکه واژگان» با دو معیار آماری تلفیق شده است (بخش ۳-۱۳-۱) و نتایج شباهت را در زمینه ارزیابی ربط بین واژگان و ربط بین فعل‌ها بهبود داده است.

Reza Javadzadeh, Morteza Zahedi, Marziea Rahimi. ”Improving verb similarity by incorporating statistics to WordNet-based metrics”. under-review.

۳. در این روش چند توصیف‌گر که در بخش ۳-۱۴-۱ توضیح داده شده‌اند، باهمدیگر تلفیق شده و با معماری شبکه عصبی Siamese برای ارزیابی شباهت زوج‌جمله استفاده شده‌اند.

Reza Javadzadeh, Morteza Zahedi, Marziea Rahimi. ”Multi-aspect embeddings and parallel training improve sentence similarity assessment”. under review

فهرست مطالب

فهرست جداول

فهرست تصاویر

لیست واژگان معادل

لیست علائم

X بردار مقادیر هدف ۱۱۴

Y بردار مقادیر پیش‌بینی شده ۱۱۴

α پارامتر قابل تنظیم ۳۶

$\bar{X}$ میانگین عناصر بردار مقادیر هدف ۱۱۴

$\bar{Y}$ میانگین عناصر بردار مقادیر پیش‌بینی شده ۱۱۴

β پارامتر قابل تنظیم ۳۶

δ پارامتر قابل تنظیم ۷۷

θ پارامتر قابل تنظیم ۴۰

ζ پارامتر قابل تنظیم ۴۰

f^t بسامد رخداد کل ۸۱

s_1 جمله اول ۳۸

s_2 جمله دوم ۳۸

LCS نزدیک‌ترین مفهوم مشترک ۳۶

depth عمق، فاصله تا ریشه ساختار درختی ۳۶

sim تابع ارزیابی شباهت ۳۵

فصل ۱: مقدمه

طی دو قرن گذشته، بشریت با موفقیت زمینه خودکارسازی بسیاری از کارها را با استفاده از دستگاه‌های مکانیکی و الکتریکی فراهم نموده است و این دستگاه‌ها به‌خوبی در زندگی روزمره به انسان خدمت می‌کنند. در نیمه دوم قرن بیستم بود که توجه انسان به خودکارسازی پردازش زبان طبیعی معطوف شد. اکنون مردم نه‌تنها درزمینه مکانیکی، بلکه در زمینه‌های فکری نیز خواهان کمک ماشین‌ها هستند. ازجمله اینکه دستگاه بتواند متن غیر آماده را بخواند، درستی آن را بیازماید، دستورالعمل‌های موجود در متن را اجرا کند یا حتی آن را به‌اندازه‌ای درک کند که بر اساس معنی آن پاسخ معقولی ارائه دهد البته انسان‌ها می‌خواهند وظیفه تصمیم‌گیری نهایی را خود بر عهده داشته باشند [۱۷].

ضرورت پردازش متن به‌طور عمده از دو موقعیت زیر ناشی می‌شود که هر دو با حجم متن تولیدشده و مورداستفاده امروز در جهان، مرتبط هستند:

- بعضی در سراسر جهان که با متون سروکار دارند، از دانش و تحصیلات تخصصی برخوردار نیستند. به‌عنوان مثال، یک منشی دفتر نمی‌تواند هر بار صدها قانون مختلف لازم را برای نوشتن نامه تجاری خوب به یک شرکت دیگر در نظر بگیرد، به‌ویژه اگر که به زبان مادری او نباشد.
- در بسیاری از موارد، برای تصمیم‌گیری کاملاً آگاهانه در مورد یک موضوع و یا یافتن اطلاعات، باید تعدادی متن را بخوانید، درک کنید و در نظر بگیرید که ممکن است هزاران برابر بیشتر از آنی باشد که یک فرد ازنظر توانی قادر به خواندن آن در طول کل زندگی خود می‌باشد.

بنابراین پردازش زبان طبیعی به یکی از اصلی‌ترین مباحث برای تبادل اطلاعات تبدیل‌شده است. توسعه سریع رایانه‌ها در دو دهه گذشته امکان اجرای بسیاری از ایده‌ها برای حل مشکلاتی که حتی تصور نمی‌شد آن‌ها به‌طور خودکار حل شوند، فراهم کرده است.

پردازش هوشمند زبان طبیعی مبتنی بر علمی به نام زبان شناسی محاسباتی است. زبان شناسی محاسباتی با زبان شناسی کاربردی و به‌طورکلی با علوم زبان شناسی ارتباط تنگاتنگی دارد.

۱-۱ دیدگاه کلی

رشد روزافزون کاربران و تولید محتوا در بسیاری از شبکه‌های اجتماعی و پلتفرم‌های تولید محتوا نیاز مبرمی برای تولید سیستم‌های خلاصه‌سازی خودکار متون ایجاد می‌کند. روش‌های ارزیابی شباهت می‌توانند از چند جمله شبیه به هم یکی را انتخاب کنند و بدین ترتیب خلاصه‌سازی خودکار انجام دهند. همچنین با انتخاب پاسخ متناظر به شبیه‌ترین جمله پرسشی، سیستم‌های ارزیابی شباهت می‌تواند نقش مهمی درزمینه طراحی سیستم‌های پرسش و پاسخ خودکار داشته باشند. تشخیص محتواهای دوباره‌نویسی شده در مقاله‌های عمومی و علمی نیز مسئله بسیار مهمی است که توسط سیستم‌های ارزیابی شباهت می‌تواند مورد کاوش قرار گیرد. درزمینه ارزیابی احساسات و یا رده‌بندی نظرات، روش‌های محاسبه شباهت می‌توانند ارزیابی اولیه مناسبی فراهم کنند. برای حل مسئله شباهت جمله، ابتدا باید روشی برای ارزیابی شباهت زوج‌واژگان محاسبه کرد. بدین منظور ابتدا روش‌های ارزیابی شباهت زوج‌واژگان بررسی می‌شود، سپس روش‌های تعمیم مسئله به ارزیابی شباهت زوج‌واژه بحث و بررسی می‌شود. همچنین وقتی شبکه عمیق برای یادگیری چند کلاس ثابت آموزش داده می‌شود، اگر بخواهیم نمونه کلاس جدیدی ایجاد کنیم ممکن است نیازمند به‌روزرسانی وزن‌های کل شبکه باشیم اما اگر به‌جای این کار شباهت داده جدید (تصویر یا متن) را با داده‌های موجود در دسته‌های آموزش مقایسه کنیم می‌توانیم با پیشنهاد یک معیار ارزیابی شباهت، کلاس شبیه‌ترین داده آموزشی را به داده جدید اختصاص دهیم [۱۲۰]. یک‌زبان مانند زبان انگلیسی پتانسیل قدرتمندی برای برقراری ارتباط بین جوامع مختلف دارد که با استفاده از آن می‌توان یک مسئله را به شکل‌های مختلفی بیان کرد. البته همین آزادی استفاده از واژگان، باعث می‌شود تحلیل گفتار برای ماشین مشکل باشد و در نتیجه، معنی برگرفته از زبان به دلیل وجود تعداد نامحدود جمله، مشکل است.

بااینکه ساز و کار درک معنا در ذهن انسان هنوز به‌عنوان یک راز باقی‌مانده است، اما گفته می‌شود که شباهت بین دو رخداد یا شیء مبتنی بر روشی است که انسان آن‌ها را توصیف می‌کند [۴۳].

اما این فرضیه که: واژگان شبیه به هم معمولاً در متون «مشابه» باهم دیده می‌شوند [۴۶]، همیشه

درست نیست. زیرا که ممکن است واژگانی مانند تکانه، نویز و بو در تناقض با واژگانی باشند که برای توصیف یک فرد به کار می‌رود مثل او جوان/زیبا/فعال است.

به فرآیند ماشینی کردن مهارت‌های عمومی و تخصصی زبان، زبان‌شناسی رایانشی گفته می‌شود. مهارت‌های عمومی زبان شامل مواردی چون تولید و درک گفتار است که نیازی به آموزش رسمی ندارند و مهارت‌های تخصصی نیز شامل مواردی چون خواندن، نوشتن و ترجمه کردن است که نیاز به آموزش دارند. اهمیت ماشینی کردن این مهارت‌های عمومی و تخصصی زبان نیز رفاهی است که به دنبال فتّاوری حاصل از آن برای بشر به وجود می‌آید.

امروزه زبان شناسان رایانشی به‌عنوان اعضای گروه‌های میان‌رشته‌ای به فعالیت می‌پردازند، که اعضای این گروه‌ها می‌توانند شامل زبان‌شناسان (درزمینه زبان‌شناسی همگانی تخصص دارند)، کارشناسان زبان (افرادی با پیش‌زمینه و تا حدی دارای مهارت‌های عملی مرتبط با پروژه موردنظر)، و دانشمندان علم کامپیوتر باشند. به‌طورکلی، زبان‌شناسی رایانشی از همکاری دانشمندان و کارشناسان رشته‌های زبان‌شناسی، علوم رایانه‌ای، متخصصین زمینه هوش مصنوعی، ریاضی، منطق، علوم شناختی، روان‌شناسی شناختی، روان-زبان‌شناسی، مردم‌شناسی، عصب‌شناسی و برخی دیگر رشته‌ها تشکیل می‌شود.

مدل‌های توزیعی و معنایی واژگان مبتنی بر اطلاعات آماری هم‌رخدادی واژگان در مجموعه بزرگی از متون هستند [۸۵]. چالش اصلی استفاده بهینه از این اطلاعات آماری است. برای مثال بااینکه واژه "و" در کنار واژه "خیابان" ممکن است دیده شود اما توصیفگر واژه خیابان نیست و باید در تشکیل بردار توصیفی خیابان مورد استفاده قرار نگیرد.

ارزیابی شباهت با استفاده از عبارت‌های متشکل از چند واژه، از حوزه‌های پژوهشی فعال امروزه است. اما در مقایسه با مدل‌های توزیعی دیگر که داده برای آن‌ها زیاد است، در این حوزه مشکل پراکندگی داده یا کمبود داده وجود دارد [۹۷]. زیرا عبارت‌های چند واژه، تنوع زیادی دارند ولی امکان رخداد هرکدام می‌تواند بسیار کم باشد. حتی اگر ما تمام کتاب‌های روی زمین را جمع‌آوری کنیم، ممکن است بعضی جمله‌ها حتی فقط یک بار تکرار شوند [۷۹].

مدل‌های ترکیبی (توزیعی) یادشده به دنبال یافتن عملگری برای ترکیب واژه‌ها هستند چرا که ممکن

است بین واژه‌ها مانند آنچه در درخت‌های وابستگی جمله‌ها می‌بینیم ارتباط وجود داشته باشد. بر طبق این مدل‌های توزیعی، معیار اندازه‌گیری شباهت یافتن یک طرح وزن‌دهی برای دو توصیف تشکیل شده جمله است. ابتدا اطلاعات مفید هم‌رخدادی از اطلاعات اضافه متمایز می‌شوند، سپس طرح وزن‌دهی، تأثیر نوع مخرب را کمینه می‌کند. در مدل‌های توزیعی فرض این است که چند واژه همسایه (هم‌رخداد) معتبر وجود دارد و همه اطلاعات کارآمد نیستند.

روش‌های ابتدایی که برای محاسبه شباهت سندها طراحی شدند مبتنی بر نرخ تعداد واژگان مشترک در دو سند بودند. اما در محاسبه شباهت بین دو جمله (یا متن کوتاه)، اطلاعات موجود بسیار محدود است. برای واژگان جمله باید ابهام‌زدایی صورت گیرد. بعضی واژه‌ها باید باهمدیگر تشکیل یک عبارت دهند تا معنا داشته باشند. این موضوع را هم به این مسئله اضافه کنید که فقط در توییتر روزانه حدود ۱۰۰ میلیون پست فرستاده می‌شود. از کاربردهای مهم ارزیابی شباهت می‌توان به برنامه گفتگوگر[1] اشاره کرد. این برنامه‌ها اهمیت روزافزونی برای فروشگاه‌ها دارند چرا که مشتریان ممکن است به فروشنده‌ها اعتماد کافی نداشته باشند و به دنبال راه‌های دیگری برای خرید کردن می‌گردند. همچنین این برنامه‌ها می‌توانند هزینه‌های یک شرکت را به‌شدت کاهش دهند. این مهم، می‌تواند به‌وسیله وجود مراکز تماس با اپراتور خودکار، دستیارهای هدایتگر هواپیما یا خودروها باشد زیرا که نیاز به استخدام کارمند را به‌شدت کاهش می‌دهد. منظور از برنامه گفتگوگر، تبادل گفتگو از طریق متن است. در ارزیابی شباهت، یک الگوریتم تعیین می‌کند که دو متن چقدر به هم شبیه هستند [۷۶]. در نتیجه، این حوزه پژوهشی به‌سرعت در حال رشد است و کاربردهای وسیعی دز زمینه‌های مختلف دارد برای مثال:

- در سیستم‌های پرسش و پاسخ، می‌توان پاسخی که شبیه‌ترین به سؤال است را بازیابی کرد [۲۳].
- همچنین در سیستم‌های پرسش و پاسخ می‌توان پاسخ‌های مشابه را به ترتیب شباهت نمایش داد.
- می‌توان در طراحی موتورهای جستجو برای نمایش ترتیبی شبیه‌ترین پاسخ‌ها به ورودی کاربر از این روش‌ها استفاده کرد.
- می‌توان در خلاصه‌سازی متون به روش استخراجی جمله‌های مهم‌تر را شناسایی و به ترتیب

[1] Conversational agents

نمایش داد. درروش خلاصه‌سازی استخراجی جمله‌های مهم استخراج‌شده و جمله‌های اضافه حذف می‌شوند [۵].

- می‌توان با شناسایی واژه‌های مهم ترکیب بهتری از واژگان را برای تولید بردار ویژگی جمله انتخاب کرد [۹۰].

به‌عنوان یک راه‌حل ابتدایی می‌توان واژگان مشترک بین دو جمله را معیار ارزیابی قرار داد. اما این روش نمی‌تواند معیار خوبی باشد زیرا که دو جمله ممکن است واژگان مشترک بسیاری داشته باشند ولی به دو چیز مختلف اشاره کنند. همچنین با افزایش طول هر جمله، احتمال پیدا شدن واژگان مشترک بین آن دو بیشتر می‌شود، در حالی که ممکن است دو جمله بی‌ربط باشند.

۱-۲ واژگان

برای ارزیابی شباهت جمله‌ها ابتدا نیاز است تا واژگان را مقایسه کنیم. هرچند بیشتر این کلمه‌های انگلیسی، شناخته‌شده هستند اما معنای آن‌ها وابسته به مفاد متن است. ویژگی‌های عام واژه شامل موارد زیر است:

- ترکیب یا ریخت: اینکه واژه چگونه هجی می‌شود.
- معنا: منظور از استفاده از این واژه چیست.
- ریشه: واژه پایه که به آن پیشوندها اضافه شده است تا به صورت کنونی‌اش شکل بگیرد.
- پسوند: واژک متصل شده به واژه که در واقع به انتهای ریشه واژه می‌چسبد.
- نوع: نقش‌واژه ریشه واژه مورد نظر.

ریشه می‌تواند چندین حالت پسوند به‌طور هم‌زمان داشته باشد که هرکدام ترکیب متفاوتی از واژه را تشکیل دهند. ترکیبی از یک واژه می‌تواند بیشتر از یک معنا داشته باشد و بیشتر از یک ریشه و نوع نیز داشته باشد. در مقابل هر معنا ریشه یکتا دارد و نوع مشخص دارد.

بیایید واژه «dogs» را در نظر بگیریم، ترکیب یا ریخت آن به سادگی نحوه نوشتار آن است به صورت «s» ،«g» ،«o» ،«d» یا «dogs». ریشه واژه «dog» است بنابراین پسوند آن «s» است. ریشه برای این واژه شامل بیش از یک نوع است چراکه می‌تواند اسم، صفت یا فعل باشد. می‌تواند بیش از یک معنا داشته باشد به‌طور مثال از معنای ممکن آن «دنبال کردن» به صورت فعل، «تنبلی» به صورت صفت و «سگ» به عنوان اسم است.

۱-۳ جمله

جمله واحد پایه گرامر است و ساختار آن توسط نشانه‌گذاری‌ها مشخص است. در ارتباط با واژگان ، برای ارزیابی شباهت جمله‌ها تعریف بعضی واژگان باید مشخص باشد. هر جمله یک «موضوع» و یک «معنا» دارد:

- موضوع: موضوع جمله مسئله‌ای است که راجع به آن صحبت می‌شود. واژه کلیدی در جمله‌ها اطلاعات مهمی را در ارتباط با موضوع جمله دارا هستند.
- معنا: اندیشه و تصوراتی که در جمله منتقل می‌شود را بیان می‌کند. این شامل ایده جمله و عملی نیز هست که در جمله به آن اشاره شده است.
- تعامل واژگان: اینکه واژگان چگونه تلفیق می‌شوند تا معنایی فراتر از استفاده از چند واژه مجزا داشته باشند.
- بافت: چطور معنی واژگان با توجه به واژگان همسایگی تعریف می‌شود. این یک مسئله سخت مبتنی بر تصویر موردبحث جمله است.

۱-۴ موضوع

نیاز است تا قبل ادامه بحث بین موضوع جمله و تعاملات واژگان تمایز مشخصی قائل شویم. موضوع مشخص می‌کند که جمله درباره چه چیزی است.جمله‌ها ممکن است معانی متفاوتی داشته باشند ولی

در مورد یک موضوع باشند. برای مثال جستجوی پرسش زیر را در موتور جستجوگر در نظر بگیرید:

مریخ چقدر سرد است؟

یک موتور جستجوگر واژگان کلیدی «مریخ» و «سرد» را شناسایی می‌کند. همه پاسخ‌های یافت شده باید در مورد موضوع سرد بودن مریخ باشند و همه نتایج باید حداقل یکی از این دو کلیدواژه را دارا باشند اما لازم نیست که همه یک معنی داشته باشند [۹]. برای مثال دو پاسخ زیر توسط موتور جستجوگر گوگل را بررسی می‌کنیم.

بیشتر مناطق مریخ بسیار سرد است

لایه‌های هوای سرد کمک می‌کنند که یخ پایدار باقی بماند و مانع گرم شدن و ناپدید شدنش می‌شوند.

هر دو پاسخ پرسش هستند اما معنای متفاوتی دارند. تفاوت بسیاری بین جمله اول و دوم وجود دارد. تفاوت در کلیدواژه‌های آن‌ها، نمایانگر تفاوت موضوع دو جمله است. بااینکه هردو به پاسخ مرتبط هستند اما ترکیب متفاوت واژه‌ها و تعاملات آن‌ها دو جمله مختلف را می‌سازند. هردو به پرسش مرتبط هستند اما کلیدواژه‌های مورد استفاده در آن‌ها متفاوت است. این نشان می‌دهد که موضوع توسط کلیدواژه‌ها مشخص می‌شوند و جدا از نحوه تعامل واژگان، می‌توانند توسط این واژه‌ها تشریح شوند.

در واقع می‌توان موضوع دو جمله را با مقایسه کلیدواژه‌ها مقایسه کرد بنابراین ارزیابی شباهت موضوع جمله‌ها می‌تواند به شباهت زوج‌واژگان تقسیم شود.

۱-۵ تعامل واژگان

وقتی واژه‌های موجود در یک جمله یکسان با واژه‌های جمله دیگر باشد ممکن است دو جمله مرتبط باشند ولی معنای یکسانی نداشته باشند. زوج‌جمله زیر را مشاهده کنید که نشان می‌دهد حتی اگر موضوع دو جمله یکی باشد، ممکن است معنای متفاوتی داشته باشند. اگر واژه‌های موجود در دو جمله یکی باشند اما معنی متفاوتی داشته باشند، به دو جمله هم‌بیان[۲] گفته می‌شود. در اینجا، هر دو جمله

[2]Hymonym

راجع به یک موضوع هستند ولی معنای متفاوتی دارند.

صبوری، لازم نیست عجله کنید صبوری لازم نیست، عجله کنید

با بودن واژه‌های یکسان، نه ابهام‌زدایی واژگان و نه ارزیابی موضوع هیچ‌کدام نمی‌توانند بین این دو جمله تمایز قائل شوند. معنی جمله بیش از معنی واژگان مجزای آن است، به نحوه تعامل واژگان با همدیگر وابسته است و این موضوع فقط وابسته به ترتیب استفاده واژگان نیز نیست [۱۰۶].

۱-۶ بافت متن

اگرچه دو عامل موضوع و تعامل واژگان باهمدیگر معنا را می‌سازند، بافت جمله نیز برای مقایسه دو جمله مورد نیاز است تا مشخص شود واژگان به چه شکل استفاده شده‌اند. برای ابهام‌زدایی واژگان، باید با توجه به محتوای متن، یک معنی مشخص را برای واژه انتخاب کرد به عبارتی معنی واژه با توجه به متن آن مشخص می‌شود. بافت از هرکدام از معنا یا عنوان پیچیده‌تر است و بیان می‌کند که ایده‌ها در متن چگونه باهم ارتباط برقرار می‌کنند و نه واژگانی که به جمله انتزاع می‌بخشند.

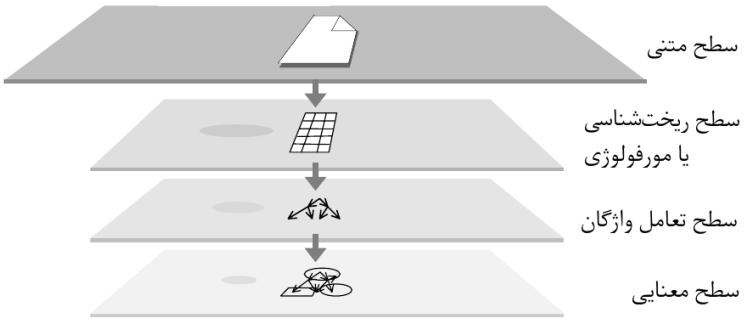
یک شخص توانایی آن را دارد که بدون در نظر گرفتن سایر جمله‌های متن، معنی یک متن را حدس بزند. بنابراین ممکن است معنی انتخاب شده توسط مطالب بعدی دستخوش تغییر شود، برای مثال

آقای سعیدی در راه با آقای مدیری برخورد کرد. صدای مشاجره آن‌ها فضا را پر کرده بود.

همان‌طور که در متن دیده می‌شود بدون دیدن جمله دوم ممکن است منظور شروع یک مکالمه باشد اما جمله دوم این معنی را می‌رساند که منظور از برخورد بحث بوده است. تصویر ۱-۱ یک توصیفگر مبتنی بر زبان‌شناسی در ارتباط با متون فراهم می‌کند.

۱-۷ شباهت

منظور از شباهت وقتی که دو چیز را شبیه هم می‌دانیم چیست؟ این پرسش برای هر مقایسه‌ای حیاتی است چراکه اجازه می‌دهد تا بتوانیم اطلاعات قبلی را به موقعیت‌های جدید تعمیم دهیم [۵۶] و این



تصویر ۱-۱: تجزیه متن به‌منظور توصیف مبتنی بر زبانشناسی [۱۷]

یکی از اصول هوشمندی می‌باشد. وقتی که بیان می‌شود دو چیز مشابه هستند، منظور آن است که ویژگی‌های مشترکی بین آن‌ها وجود دارد. اگر که بدانیم کدام ویژگی‌ها یکسان هستند آنگاه می‌توان اطلاعات راجع به یکی را دانست و به هردو تعمیم داد. این باعث می‌شود که نه تنها اطلاعات کارآمدتر ذخیره شوند، بلکه بتوان چیزهای جدید را با دانش موجود مقایسه نمود. در اصل مقایسه، مردم به دنبال شباهت بین عناصر می‌گردند. این ویژگی‌ها مدنظر ما برای محاسبه شباهت هستند [۷۳].

۱-۸ هدف

پژوهش‌های بسیاری برای درک زبان و نحوه تفسیر و تشکیل آن توسط انسان، انجام شده است. هدف از این پژوهش، جمع‌بندی روش‌های پیشین و ارائه یک روش جامع‌تر برای ارزیابی شباهت دو جمله است. به عبارتی نحوه برطرف کردن نقص‌های روش‌های پیشین در این پژوهش بررسی می‌شود.

امیدواریم تا مدل ارزیابی بهتری ارائه دهیم که بتواند درک بیشتری از تفسیر جمله ارائه دهد. همچنین تعدادی از روش‌های ارائه شده به شکل بدون ناظر هستند و نیاز به آموزش روی داده‌ها ندارند در نتیجه سریع و کارآمد هستند.

مسئله دوم که حل آن مورد بررسی قرار گرفته، مشکل ارزیابی شباهت زوج‌واژگان است هنگامیکه یک یا هردو واژه «فعل» هستند. روش‌های مبتنی بر پایگاه دانش بیشتر متشکل از اسم‌ها هستند و روش‌های آماری نیز نتوانستند به تنهایی ارزیابی خوبی داشته باشند.

ارزیابی شباهت دو جمله یک ابزار مهم برای خودکارسازی عملیات مختلف در زبان است. برای مثال

یک معیار ارزیابی شباهت می‌تواند در طراحی سیستم‌های پرسش و پاسخ مورد استفاده قرار بگیرد تا در یک کاربرد ساده، پاسخی که شبیه‌ترین به جمله پرسشی است به‌عنوان پاسخ به کاربر نمایش داده شود و در کاربرد پیچیده یک معیار ارزیابی مناسب ممکن است درک بهتری از تفسیر متن ارائه دهد [۷۲].

۱-۹ شباهت زوج‌جمله چیست؟

ارزیابی شباهت زوج‌جمله‌ها به‌سادگی می‌توان به صورت زیر تعریف شود:

معیار ارزیابی شباهت معنایی بین یک زوج‌جمله

موضوع پژوهش این رساله در مورد زبان انگلیسی است و طول دو جمله در ارزیابی متغیر اما کوتاه در نظر گرفته می‌شود. تمرکز این رساله پژوهشی در ارتباط با جمله‌هایی است که واژگان یکسان ندارند و یا فعل جزو عناصر اصلی جملات محسوب می‌شوند. زیرا که روش‌های مبتنی بر پایگاه دانش، بیشتر از نقش‌واژه اسم تشکیل شده‌اند و پوشش ضعیفی در زمینه نقش‌واژه فعل دارند.

به‌طورکلی دو روش در این بخش وجود دارد. درروش اول زوج‌واژه‌های مشترک/هم‌معنی با همدیگر منطبق می‌شوند و شباهت کل با جمع کردن شباهت اجزا به دست می‌آید که آن را روش مبتنی بر معنای واژه نام‌گذاری می‌کنیم. در روشی دیگر، شباهت دو جمله ابتدا با تشکیل یک بردار ویژگی متحد برای هر جمله و سپس محاسبه شباهت بین این دو بردار به دست می‌آید که آن را مدل معنایی نام‌گذاری می‌کنیم. بزرگ‌ترین مشکل این روش‌ها این است که وابسته به قواعد برچسب‌گذاری شده توسط انسان هستند و در ابهام‌زدایی ضعیف هستند.

روش‌های مبتنی بر شبکه واژگان هم که برای انسان‌ها و هم ماشین قابل‌درک هستند موفقیت چشمگیری در این زمینه داشتند. از مزایای این سیستم‌ها این است که روند رفع اشکال آن برای انسان ساده است. اما بیشتر واژگان موجود در این پایگاه داده‌ها که به‌عنوان یک "مفهوم" شناخته می‌شوند از نقش‌واژه اسم تشکیل شده‌اند [۱۳۱].

روش‌های مبتنی بر شبکه عصبی موفقیت چشمگیری در زمینه درک معنایی جمله دارند. این

سیستم‌ها با استفاده از واحدهای حافظه کوتاه‌مدت/بلندمدت [۳] و واحدهای همگشتی دروازه‌ای [۴]، یک شبکه عصبی طراحی می‌کنند. در این شبکه‌ها سعی بر این است تا توصیف فشرده برداری برای جمله تشکیل شود به‌طوری که تفاوت دو جمله متناظر با استفاده از یک معیار فاصله مانند اقلیدسی یا کاسینوس کمینه شود.

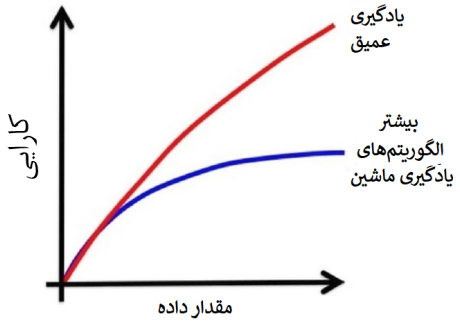
۱-۱۰ وابستگی در برابر شباهت

راه‌های دیگری هم وجود دارد که دو جمله به یکدیگر ربط داشته باشند ولی ممکن است هم‌معنی نباشند. به‌عنوان مثال می‌توان به دو فرآیند پرسش و پاسخ و ارزیابی ربط زوج‌جمله اشاره کرد که در آن به دنبال یافتن یک ارتباط منطقی بین دو جمله هستیم. درواقع، ممکن است دو جمله متضاد یا شبیه هم باشند ولی ازنظر الگوریتم ارزیابی ارتباط، نمره یکسانی دریافت کنند زیرا که اینجا ربط دو جمله، و نه نوع ربط آن، مهم است [۱۰۸]. در این پژوهش هردو حالت را بررسی می‌کنیم اما هدف اصلی مقاله ارزیابی شباهت دو جمله است.

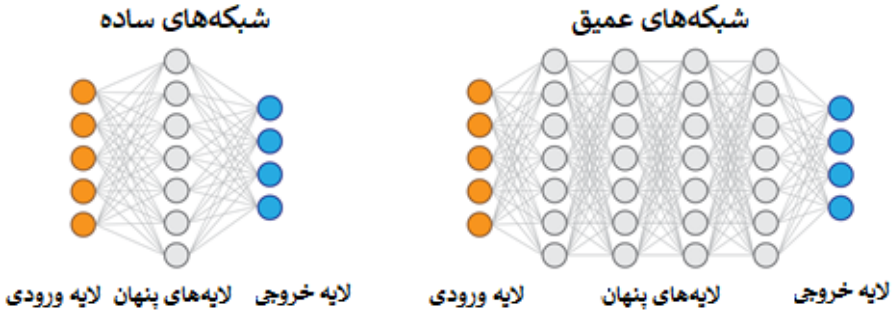
یادگیری عمیق حوزه‌ای از یادگیری ماشین است که عملکرد عالی خود را در زمینه متن، تصویر و گفتار اثبات کرده است. مجموعه الگوریتم‌های پیاده‌سازی شده تحت عنوان یادگیری عمیق شباهت بسیاری با ساختار نورون‌ها و روابط آن‌ها در مغز دارند. یادگیری عمیق کاربردهای گسترده‌ای در بینایی رایانه، ترجمه زبان، تشخیص گفتار، تولید تصاویر و موارد دیگر دارد.

اکثر الگوریتم‌های یادگیری عمیق مبتنی بر مفهوم شبکه‌های عصبی مصنوعی هستند و آموزش این الگوریتم‌ها در دنیای امروز با وجود داده‌های فراوان و منابع محاسباتی کافی، آسان‌تر شده است. همراه با داده‌های بیشتر، عملکرد مدل‌های یادگیری عمیق همچنان در حال بهبود می‌باشد. در تصویر ۱-۲ نمایش بهتری از این موضوع فراهم‌شده است. در پژوهش سان و همکاران [۱۱۶]، پژوهشگران با تحقیق روی تکنیک‌های یادگیری عمیق برای رده‌بندی ۳۰۰ میلیون تصویر دریافتند که دقت رده‌بندی با افزودن داده‌های آموزش به صورت نمایی رشد می‌کند.

[3] LSTM
[4] GRU



تصویر ۱-۲: قیاس عملکرد روش‌های یادگیری عمیق با افزایش داده‌ها [۱۱۶]



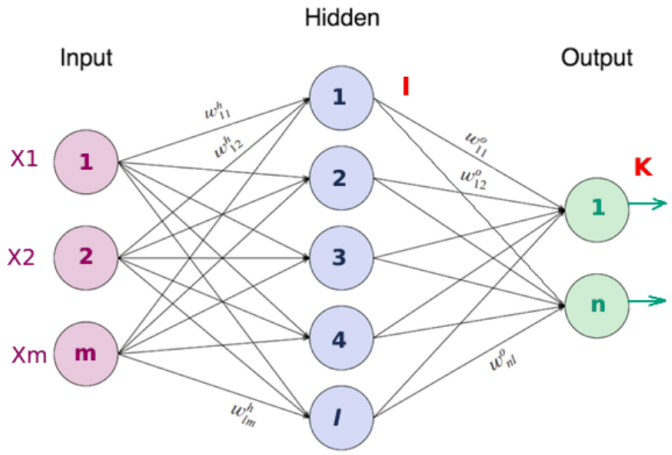
تصویر ۱-۳: قیاس شبکه‌های عصبی سنتی با عمیق [۱۳۲]

اصطلاح عمیق به عمق معماری شبکه عصبی مصنوعی اشاره دارد و یادگیری مخفف یادگیری از طریق خود شبکه عصبی مصنوعی است. تصویر ۱-۳ نمایش ساده‌ای از تفاوت بین شبکه عمیق و ابتدایی را فراهم می‌کند.

شبکه‌های عصبی عمیق قادر به کشف ساختارهای نهفته (یا یادگیری ویژگی) از داده‌های بدون برچسب و بدون ساختار هستند، مانند تصاویر (داده‌های پیکسلی)، اسناد (متنی)، یا صوتی. اگرچه یک شبکه عصبی مصنوعی و مدل‌های یادگیری عمیق اساساً ساختارهای مشابهی دارند، اما این بدان معنی نیست که ترکیبی از دو شبکه عصبی مصنوعی هنگام آموزش برای استفاده از داده‌ها، عملکردی مشابه شبکه عصبی عمیق خواهند داشت. آنچه هر شبکه عصبی عمیق را از یک شبکه عصبی مصنوعی معمولی متمایز می‌کند نحوه استفاده از مقادیر پس‌انتشار[5] است. در یک شبکه عصبی مصنوعی معمولی، الگوریتم پس‌انتشار، لایه‌های بعدی را بهتر از لایه‌های اولیه یا قبلی آموزش می‌دهد [۴۰].

برای آنکه شبکه‌های عصبی سریع‌تر و کاراتر آموزش‌دیده شوند، تعداد متنوع و زیادی از این شبکه‌ها طراحی شدند تا برای هر مسئله به بهترین شکل آموزش ببینند. این باعث شده است که شبکه‌های

[5]Backpropagation



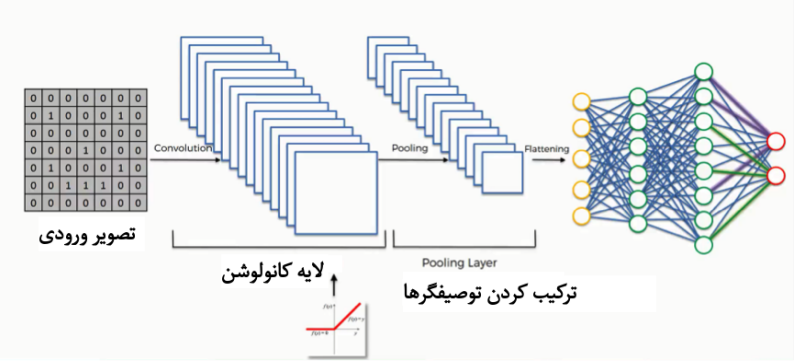
تصویر ۱-۴: نمونه شبکه پیش‌خور برای حل مسئله رده‌بندی [۵۸]

عصبی با معماری‌های متنوع که هرکدام خروجی نورون مختلفی دارند طراحی شود همچنین توابع مورد استفاده در نورون‌ها دارای تنوع است.

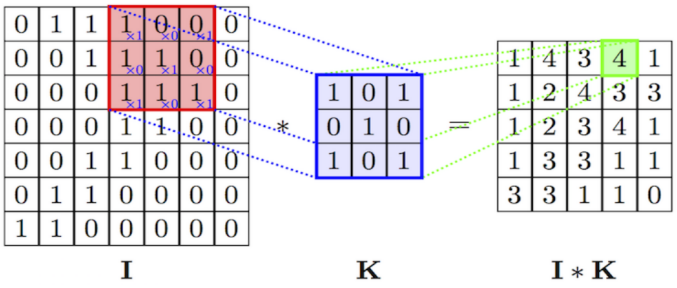
۱-۱۱ Feedforward Neural Networks

در شبکه‌های عصبی پیش‌خور[۶]، ارتباط بین نورون‌ها از لایه‌های بعدی به قبلی به‌صورت بازگشتی نیست. این نوع شبکه‌ها پایه خانواده شبکه‌های عصبی را تشکیل می‌دهند. حرکت داده‌ها از لایه ورودی به خروجی ، از طریق لایه‌های پنهان موجود است و هیچ نوع حلقه‌ای وجود ندارد. خروجی از یک لایه در ساختار شبکه به‌عنوان ورودی به لایه بعدی عمل می‌کند. همان‌طور که در تصویر ۱-۴ دیده می‌شود n نورون در انتها وجود دارند که می‌تواند به معنی معماری شبکه به منظور رده‌بندی باشد. با توجه به اینکه کدام‌یک از n کلاس هدف موردنظر است، باید وزن‌های شبکه طوری آموزش داده شود تا نورون n بیشترین مقدار خروجی را تولید کند و بقیه نورون‌ها به‌اصطلاح خاموش باشند.

[6] Feedforward



تصویر ۱-۵: نحوه عملکرد شبکه‌های پیچشی [۷۰]



تصویر ۱-۶: پیچش پنجره بر روی ورودی [۸۹]

۱-۱۲ شبکه‌های عصبی پیچشی

شبکه‌های کانولوشن یا پیچشی [۷] برای تشخیص تصویر و تشخیص دست خط عالی هستند. ساختار آن‌ها بر اساس نمونه‌برداری از یک پنجره یا بخشی از تصویر، سپس تشخیص ویژگی‌های آن و بعد استفاده از این ویژگی‌ها برای ساختن یک توصیفگر است. همان‌طور که می‌توان تخمین زد، این منجر به استفاده از چندین لایه می‌شود و این مدل‌ها اولین مدل‌های یادگیری عمیق هستند. به تصویر ۱-۵ نگاه کنید.

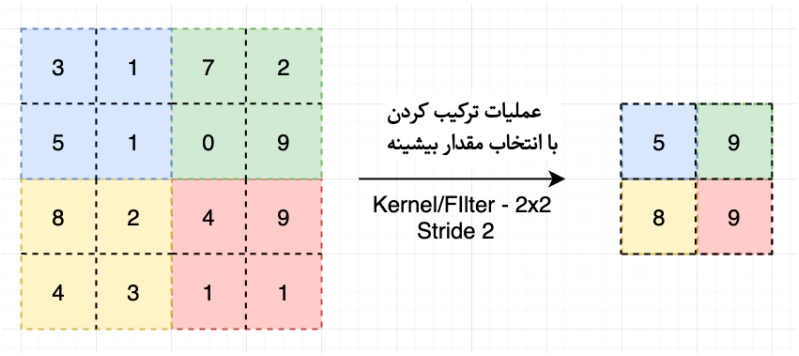
همان‌طور که در تصویر مشخص است، چندین لایه پیچشی و تجمعی [۸] وجود دارد. در مرحله پیچش مانند تصویر ۱-۶، تعدادی پنجره با اندازه‌های از پیش مشخص‌شده بر روی تصویر ورودی اعمال می‌شوند. درنتیجه تصویر (یا توصیف) جدیدی خواهیم داشت که باید ویژگی‌های مهم تصویر قبلی را استخراج کرده باشد.

در مرحله بعد مانند تصویر ۱-۷، ویژگی‌های برجسته‌تر استخراج می‌شوند به این معنی که برای مثال

[7] Convolutional

[8] Pooling

میانگین مقادیر در یک پنجره، یا حداکثر مقدار آن‌ها استخراج می‌شوند، عملگرهای دیگری نیز می‌توان تعریف کرد اما به‌طور معمول از موارد یاد شده استفاده می‌شود.



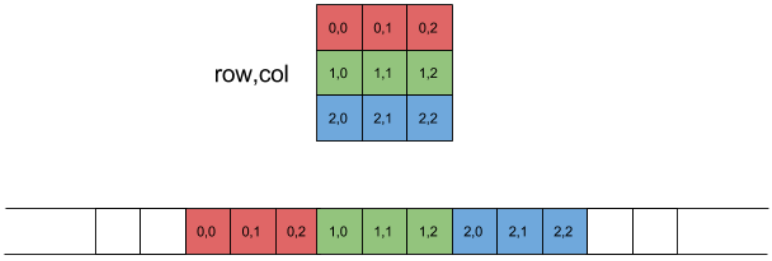
تصویر ۱-۷: عملیات تجمعی بر روی توصیفگر [۶۷]

در مرحله نهایی به‌طور معمول ابتدا مانند تصویر ۱-۸ ماتریس توصیفگر دوبعدی به یک بردار تبدیل‌شده و سپس به یک شبکه عصبی پیش‌خور خورانده می‌شود.

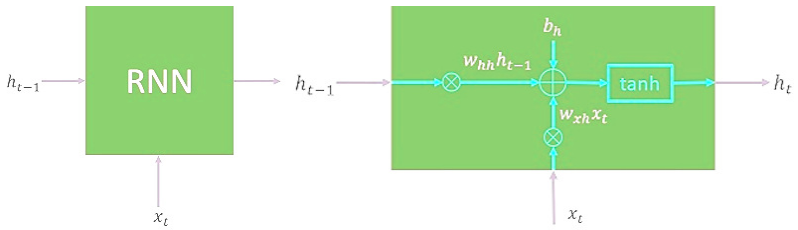
RecNN ۱۳-۱

نوع شبکه‌های عصبی همگشتی هنگامی که الگوها در طول زمان تغییر می‌کنند مورد استفاده قرار می‌گیرند و درواقع در طی زمان الگوهای داده‌ها را پیدا می‌کنند. شبکه عصبی همگشتی نوعی شبکه عصبی است که در آن یک ارتباط بین گره‌ها طی یک توالی زمانی وجود دارد. این ارتباط ، یک گراف جهت‌دار است. منظور از توالی زمانی ، داده‌هایی است که با گذر زمان انتقال می‌یابند [۱].

مثال RNN: داده‌های سری زمانی ، شامل قیمت سهام‌ها که با تغییر زمان ، اطلاعات ثبت‌شده از حسگرها و سوابق پزشکی متغیر هستند. این شبکه‌ها از حالت داخلی یا حافظه خود برای پردازش



تصویر ۱-۸: تخت کردن توصیفگر دوبعدی [۱۲]



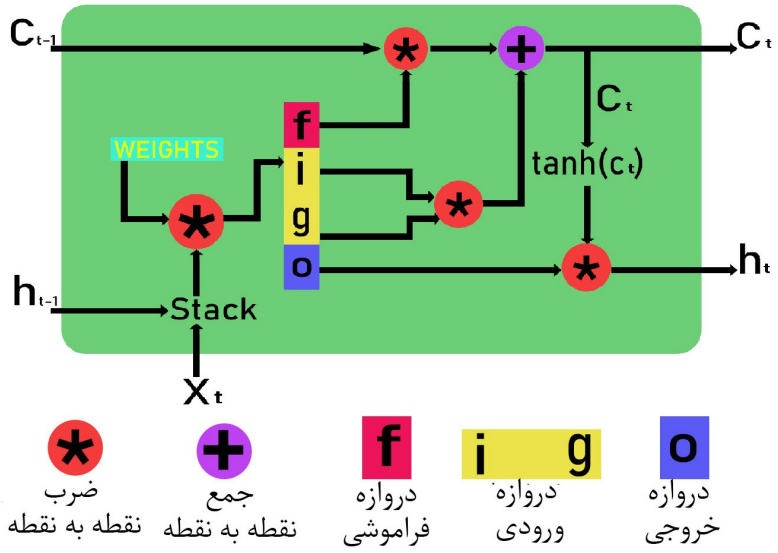
تصویر ۱-۹: شبکه‌های بازگشتی RNN تصویر الهام گرفته شده از [۱۴۲]

دنباله داده‌های ورودی استفاده می‌کنند. این قبیل ورودی‌ها به ورودی قبلی وابستگی دارند. بنابراین، یک ارتباط بین دنباله‌های ورودی وجود دارد. لذا از آن‌ها در حوزه‌هایی مانند پردازش زبان طبیعی و تشخیص گفتار استفاده می‌شود. شبکه RNN، LSTM و GRU سه نوع از پرکاربردترین معماری‌های شبکه بازگشتی هستند. نحوه کار شبکه‌های RNN به‌صورت تصویر ۱-۹ است و از فرمول ۱-۱ محاسبه می‌شود [۱۱۷].

$$h_t = tanh(w_{hh}h_{t-1} + w_{xh}x_t + b_h) \qquad (۱-۱)$$

همان‌طور که در تصویر مشخص است خروجی هر دروازه RNN مساوی با رابطه جمع اطلاعات قبلی و ورودی جاری است. این بدان معنی است که شبکه باید در طول آموزش، یاد بگیرد که اطلاعاتی از گذشته را که در تصمیم‌گیری نهایی مهم است یا یاد بسپارد. این مسئله به وضوح در رابطه ۱-۱ مشخص است. همان‌طور که در تصویر مشخص است به تعداد ورودی‌های زمانی از RNN استفاده می‌کنیم. درواقع RNNهای نمایش داده شده همه یکی هستند و به ترتیب زمان نمایش داده شده‌اند. اما از مشکلات این شبکه‌ها می‌توان موارد زیر را اشاره کرد.

- به خاطر طبیعت بازگشتی بودن مقادیر، محاسبات کند است.
- آموزش آن مشکل است.
- مشکل تضعیف مقادیر محاسبه شده گرادیان در طی زمان وجود دارد بنابراین آموزش آن مشکل است.



تصویر ۱-۱۰: نمایی از واحد حافظه LSTM [۱۱۱]

- در عمل توانایی به یادسپاری خوبی ندارد.

- وقتی از یک تابع غیرخطی برای نورون‌ها استفاده کنیم دنباله‌های طولانی را نمی‌شود به‌خوبی پردازش کرد.

۱-۱۴ حافظه‌های کوتاه‌مدت و بلندمدت

این نوع شبکه‌ها[9] به عبارتی LSTM حافظه‌هایی هستند که توانایی به خاطر آوردن اطلاعات مهم چندین گام قبل و گام‌های اخیر را دارا هستند [۵۰]. می‌توان گفت وقتی که از LSTM به جای RNN استفاده شود، درواقع پیچ‌های بیشتری برای تنظیم کردن ورودی‌ها داریم و درنتیجه نحوه تلفیق ورودی‌ها با وزن‌های آموزش‌دیده شده بهتر کنترل می‌شوند. و بنابراین انعطاف‌پذیری بیشتری در تولید خروجی داریم هرچند که پیچیدگی محاسباتی و عملیاتی آن بیشتر است. نمونه این شبکه‌ها را در تصویر ۱-۱۰ مشاهده کنید.

داشتن یک حافظه بلندمدت‌تر دارای اثر تثبیت‌کننده است چراکه حتی اگر شبکه نتواند از تاریخچه اخیر خود درک صحیحی پیدا کند، باز بااین‌وجود قادر است با نگاه در گذشته پیش‌بینی خود را کامل

[9] Long short-term memory

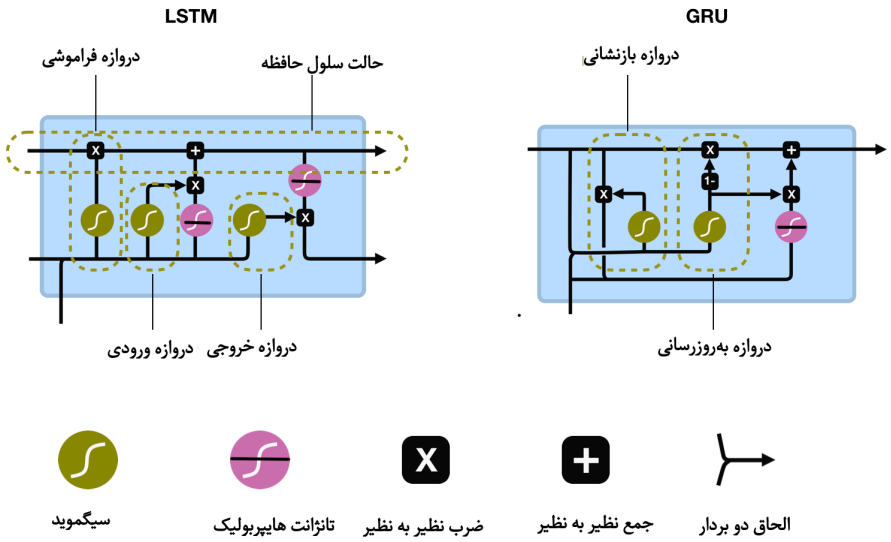
کند. برخلاف شبکه عصبی بازگشتی سنتی که در آن محتوا در هر گام زمانی از نو بازنویسی می‌شود، در یک شبکه عصبی بازگشتی LSTM شبکه قادر است نسبت به حفظ حافظه جاری از طریق دروازه‌های معرفی شده تصمیم‌گیری کند. به‌طور شهودی اگر واحد LSTM ویژگی مهمی در دنباله ورودی در گام‌های ابتدایی را تشخیص دهد به سادگی می‌تواند این اطلاعات را طی مسیری طولانی منتقل کند بنابراین این‌گونه وابستگی‌های بلندمدت احتمالی را دریافت و حفظ می‌کند. همان‌طور که در تصویر ۱-۱۱ آورده شده است دروازه فراموشی وظیفه دارد تا بفهمد با توجه به وضعیت سلول حافظه قبلی و اطلاعات جدید، کدام اطلاعات از ورودی را نگه دارد و کدام‌ها را حذف نماید.

۱-۱۵ واحد بازگشتی دروازه‌ای

شبکه‌های GRU[10] از LSTM برای حل بعضی مسائل عملکرد بهتری نشان می‌دهند. در تصویر ۱-۱۱ نمونه‌ای از حافظه GRU و مقایسه کارکرد آن با LSTM را مشاهده کنید. مقادیرGRUها آسان‌تر اصلاح می‌شود و نیازی به واحد حافظه مستقل ندارد و بنابراین سریع‌تر از LSTM آموزش داده می‌شود در حالی که کارایی آن به‌خوبی LSTM است. برای حل مشکل محوشدگی گرادیان در شبکه عصبی سنتی یکی از راه‌حل‌ها، استفاده از GRU می‌باشد. این نوع معماری از مفهوم‌هایی به نام دروازه به‌روزرسانی و بازنشانی (Update gate و Reset gate) استفاده می‌کند. این دو دروازه یا gate در اصل بسیار مشابه هستند که با استفاده از آن‌ها تصمیم گرفته می‌شود چه اطلاعاتی به خروجی منتقل شود و چه اطلاعاتی منتقل نشود. نکته خاص درباره این دروازه‌ها آن است که می‌توان آن‌ها را آموزش داد تا اطلاعات مربوط به گام‌های زمانی بسیار قبل را بدون آنکه در حین گذر زمان (طی گام ای زمانی مختلف) دستخوش تغییر شوند حفظ کند. دروازه ورودی تصمیم می‌گیرد با توجه به اطلاعات حافظه پنهان قبلی و جدید کدام باید به حافظه بلندمدت سپرده شوند. دروازه خروجی تصمیم می‌گیرد کدام اطلاعات حافظه پنهان جاری برای انتقال به بازه زمانی بعدی حفظ شود.

در حافظه نوع GRU دو نوع عملیات اصلی وجود دارد. دروازه به‌روزرسانی به مدل کمک می‌کند

[10]Gated recurrent unit



تصویر ۱-۱۱: عملیات تعریف‌شده بر دو حافظه LSTM و GRU [۱۰۳]

تا تعیین کند چقدر از اطلاعات گذشته به آینده منتقل شود. دروازه بازنشانی تصمیم می‌گیرد که چه مقدار اطلاعات گذشته فراموش شوند. نحوه محاسبه آن مانند دروازه به‌روزرسانی است اما تفاوت آن در وزن‌ها و نحوه استفاده از این دروازه است. به‌طورکلی GRU دروازه‌های محاسباتی کمتری دارد، سریع‌تر آموزش داده می‌شود، در مواردی بهتر عمل می‌کند و درک عملکرد آن ساده‌تر است.

۱-۱۶ نوآوری

روش‌های پیشین را از سه دیدگاه بررسی کرده‌ایم: روش‌های آماری مبتنی بر اطلاعات متون، روش‌های مبتنی بر پایگاه دانش لغوی و روش‌های توصیف‌گر مبتنی بر مدل‌های عصبی.

در روش‌های آماری، هم‌رخدادی تمامی زوج‌واژگان دو جمله در یک پیکره بزرگ محاسبه می‌شود. از روی اطلاعات به‌دست‌آمده ماتریس شباهت تمامی زوج‌واژگان تشکیل می‌شود. سپس از اطلاعات این ماتریس برای محاسبه شباهت استفاده می‌شود. برای مثال می‌توان بردار ویژگی مستقل برای هر جمله تشکیل داد و فاصله دو بردار رو محاسبه کرد [۶۱]. یا می‌توان از روش‌هایی مانند تحلیل معنایی نهفته [11] استفاده کرد [۳۰]. روش‌های مبتنی بر یادگیری عمیق ابتدا یک بردار ویژگی برای هر واژه تشکیل

[11] LSA

می‌دهند برای مثال طوری که بتوان با استفاده از بردار ویژگی واژگان همسایه، واژه مرکزی را پیش‌بینی کرد. سپس بردارهای هر واژه به ترتیب موقعیت آن در جمله به یک حافظه از نوع بازگشتی[12] خورانده می‌شوند. این حافظه وظیفه به یادسپاری اطلاعات مهم و تشکیل یک بردار ویژگی واحد برای جمله را بر عهده دارد. شباهت بردار ویژگی دو جمله،معیار شباهت آن است [۹۰].

اصلی‌ترین پایگاه دانش مورد استفاده برای ارزیابی شباهت دو جمله را، پایگاه داده لغوی WordNet در نظر می‌گیریم که حاوی اطلاعات ارتباط واژگان به‌صورت درختی است [۸۷]. پایگاه دانش مورد استفاده دیگر که مبتنی بر گراف ارتباطی واژگان در ویکی‌پدیا است Wikipedia2Vec نام دارد [۱۳۸]. پایگاه دانش YagoNet افزونه‌ای بر داده‌های WordNet است [۱۱۵]. این پایگاه دانش حاوی اطلاعات دو پایگاه دانش اشاره‌شده به همراه اطلاعات موجودیت‌های مختلف و ارتباط بین آن‌ها می‌باشد. در این روش، هر واژه جمله به‌صورت مستقل با واژگان جمله دیگر مقایسه می‌شود و تلفیقی از نمرات به‌دست‌آمده برای استخراج نمره نهایی استفاده می‌شود. از روش‌های مبتنی بر این فن که شناخته‌شده‌اند، می‌توان به موارد زیر اشاره کرد:

- STASIS که فقط اسامی موجود در WordNet و شباهت ترتیب واژگان بین دو جمله را برای محاسبه شباهت در نظر می‌گیرد [۷۶].
- DTW[13] که از الگوریتم پیچش زمانی پویا که در اصل برای محاسبه شباهت دو سیگنال مورد استفاده است، برای محاسبه شباهت دو جمله استفاده می‌کند[۱۳].
- OMIOTIS و پیچش زمانی پویا تمامی پیوندهای موجود در پایگاه دانش را برای محاسبه شباهت بین دو جمله در نظر می‌گیرند [۱۲۴].
- SymSS پیوندهای درخت پویش[14] را علاوه بر موارد بالا در نظر می‌گیرد [۹۵].

از بزرگ‌ترین مشکلات این روش‌ها می‌توان به این اشاره کرد که هنگامی واژه‌ای در پایگاه دانش مورد

[12] Recurrent

[13] Dynamic time warping

[14] Parse tree

استفاده آن‌ها نباشد، نمی‌توانند خروجی تولید کنند.

در روش‌های مبتنی بر تولید توصیف‌گر برای واژگان سعی می‌شود که با توجه به واژگان همسایگی یک توصیف از واژه مرکزی تشکیل شود. به عبارتی شبکه عصبی سعی می‌کند تا با استفاده از واژگان همسایگی، واژه هدف را تخمین بزند. از مشکلات این روش پیچیدگی محاسباتی بالا، گاهی عملکرد ضعیف در حل مسائل خاص و امکان بالای بیش‌برازش در مرحله بهینه‌سازی توصیف‌گرها برای حل مسئله خاص است.

هدف این رساله الگو گرفتن از هردو روش آماری و پایگاه دانش برای طراحی یک سیستم تلفیقی است که بتواند در صورت عدم حضور واژگان در پایگاه دانش نمره ارزیابی صحیح تولید کند و همچنین با استفاده از روش‌های پایگاه دانش بتواند نمره همبستگی قابل‌اعتمادتری محاسبه کند. همچنین نتایج روش‌های پیشین را در ارتباط با ارزیابی شباهت زوج‌فعل‌ها بهبود دهد.

از محدودیت‌های موجود، می‌توان به عدم پشتیبانی کافی پایگاه‌های دانش از نقش واژه‌های غیراسمی اشاره کرد. در مواردی که اطلاعات واژه‌ای توسط نیروی انسانی در پایگاه دانش نوشته نشده باشد امکان تولید نمره ارزیابی وجود ندارد. روش‌های ساده آماری هرچند اطلاعات هم‌رخدادی را در نظر می‌گیرند اما وابستگی بین واژگان را حساب نمی‌کنند. شبکه‌های عصبی در بیشتر موارد نیاز به بردار ویژگی پیش‌آموزش‌دیده دارند تا آموزش تشکیل بردار ویژگی جمله صورت گیرد. برای حل این مسئله پژوهش‌هایی طراحی بردار ویژگی را مبتنی بر کاراکترهای واژگان مدنظر قرار داده‌اند. هرچند این روش می‌تواند بردار ویژگی برای هر واژه‌ای تولید کند اما هزینه آن، کیفیت پایین‌تر بردار خواهد بود [۱۳۳].

چالش دیگر، نحوه تلفیق این دو روش است. روش نظارت‌شده که در آن باید ضریب‌های مناسب برای هر روش را محاسبه کرد و روش بدون ناظر که میانگین نمرات را محاسبه می‌کنیم. طبق آزمایش‌های اولیه، در تلفیق روش‌های آماری و پایگاه دانش، بهبودی با محاسبه ضرایب متفاوت حاصل نشد.

پژوهش‌هایی که در این رساله صورت گرفته است اهمیت بسزایی در پژوهش‌های مرتبط با شباهت روج‌جمله‌ها دارد. نحوه استفاده از هردو روش آماری و پایگاه دانش تاکنون بررسی نشده است. بردارهای ویژگی تولید شده توسط روش‌های غیر Glove و Word2Vec نیز به‌ندرت برای ارزیابی شباهت زوج‌جمله

آنالیز شده است.

دستاوردهای این پژوهش را می‌توان به موارد زیر خلاصه کرد:

- طراحی سیستم‌هایی که با استفاده از دانش آماری، ارزیابی شباهت زوج‌واژگان و زوج‌جمله‌ها را بهبود می‌بخشند.
- یک سیستم شبکه عصبی با نلفیق توصیف‌گرهای مکمل همدیگر طراحی شده است که عملکرد بهتری نسبت به Word2Vec و مدل RoBERTa-base کسب کرد.
- در آزمون‌ها اثبات می‌شود که روش پیشنهادی ارزیابی زوج‌واژگان در پایگاه داده‌های حاوی نقش‌واژه فعل، عملکرد بهتری دارد.
- استخراج ویژگی‌های عمومی و محلی از ماتریس شباهت زوج‌واژگان باعث بهبود نتایج در روش پیشنهادی GLP شده است.
- عملکرد توصیف‌گر ViCo که مبتنی بر هم‌رخدادی اشیاء در تصویر بود از توصیف‌گر Word2Vec روی داده‌های ارزیابی زوج‌جمله‌ها بهتر است.
- عملکرد توصیف‌گر مبتنی بر وابستگی اجزای جمله GCN به صورت میانگین از توصیف‌گرهای مستقل از متن مورد مقایسه بهتر بود. همچنین با توجه به نتایج روش وابسته به محتوای StructBERT می‌توان به این نتیجه رسید که در نظر گرفتن نحوه وابستگی واژگان اهمیت بالایی در ارزیابی شباهت جمله‌ها دارد.
- نتایج ضعیف‌تر توصیف‌گرهای مستقل از متن مانند Word2Vec روی داده‌های Stsbenchmark حکایت از نیاز بهینه‌سازی توصیف‌گرها روی مجموعه داده‌های متنوع از جمله داده‌های خبری، محتوای توییتر، تالارهای پرسش و پاسخ دارد.

روش موجک

در روش پیشنهادی مبتنی بر توصیف موجک، ابتدا روی واژگان داده‌های آموزش ماسک‌گذاری انجام شده است و سپس، توصیف‌گر (در اینجا RoBERTa-base) برای پیش‌بینی واژه حذف شده، بهینه می‌شود. سپس روی توصیف جمله تبدیل موجک انجام می‌شود. توسط تبدیل موجک، به توصیف جمله به عنوان سیگنال نگاه می‌شود و اطلاعات زمانی محلی با جزئیات بیشتر در کانال‌های جدا استخراج می‌شود. در مرحله بعد شباهت کانال‌های متناظر بین زوج‌جمله به عنوان ویژگی به رگرسیون ارسال می‌شود. حال می‌توان یک معادله خطی/غیرخطی برای محاسبه شباهت بین دو جمله محاسبه کرد. طبق آخرین اطلاعات ما، تاکنون در کارهای گذشته این رویکرد در زمینه توصیف جمله‌ها مورد استفاده قرار نگرفته است.

۱-۱۷ پرسش‌های پژوهشی

- آیا می‌توان با استفاده از داده‌های آماری بسامد رخداد روش مناسبی برای ارزیابی شباهت محاسبه کرد؟
- آیا در نبود یک مفهوم در پایگاه دانش از پیش آماده شده، می‌توانند توسط روش‌های آماری جایگزین شوند؟
- آیا با تلفیق این دو روش می‌توان روش بهتری طراحی کرد؟
- آیا شبکه‌های عصبی در ارزیابی شباهت بهتر از روش‌های مبتنی بر شبکه واژگان هستند؟
- در کدام مجموعه‌های داده، روش‌های صرفاً آماری پیش هستند؟

۱-۱۸ چالش‌ها

در محاسبه شباهت بین دو جمله، با توجه به کوتاه بودن طول هر جمله و کم بودن تعداد واژگان مورد استفاده، ارزیابی شباهت مشکل‌تر از ارزیابی شباهت بین دو سند است. در ارتباط با اسناد تعداد واژگان

مورد استفاده بسیاری وجود دارد، ولی در ارتباط با دو جمله این‌طور نیست. شباهت واژه به واژه به تنهایی نمی‌تواند نتیجه ارزیابی خوبی داشته باشد؛ ممکن است یک جمله فقط با استفاده از یک واژه منفی کاملاً نقض شود در صورتی که واژگان مشترک بین دو جمله وجود داشته باشد، درنتیجه کار ارزیابی مشکل‌تر می‌شود. به دلیل نبود اطلاعات کافی در یک جمله طراحی یک مدل آماری شرطی برای محاسبه شباهت بسیار مشکل است. روش‌های شبکه‌های عمیق در محاسبه شباهت بین دو جمله موفق هستند اما این روش‌ها در تشخیص اصطلاح‌های مورد استفاده در مقایسه با انسان کارایی کافی ندارند. برای مثال الگوریتم شبکه عصبی هنوز نمی‌تواند برای جمله «یک دست صدا ندارد» بردار ویژگی درستی تشکیل دهد و صرفاً بردار ویژگی هرواژه را به ترتیب کدنگاری می‌کند تا بردار ویژگی جمله را تشکیل دهد. این در صورتی است که معنی جمله یادشده ممکن است «موفقیت‌های بزرگ با کمک و همکاری همه به دست می‌آید» باشد.

شبکه واژگان شامل لیست وسیعی از مفاهیم و ارتباط بین آن‌ها به‌صورت درختی است؛ به‌عنوان مثال در تصویر زیر نمونه‌ای آورده شده است. در تصویر ۱-۱۲ نحوه ارتباط نام‌ها آورده شده است. معیارهای شباهت مبتنی بر این شبکه، فاصله بین هر واژه (مفهوم) با نزدیک‌ترین مفهوم مشترک زوج‌واژه، عمق مفهوم مشترک نسبت به ریشه شبکه، اطلاعات بسامد هر واژه (محتوای اطلاعات) را برای محاسبه شباهت در نظر می‌گیرند. در فصل بعد نمونه‌ای از این معیارها و نحوه محاسبه شباهت آورده شده است.

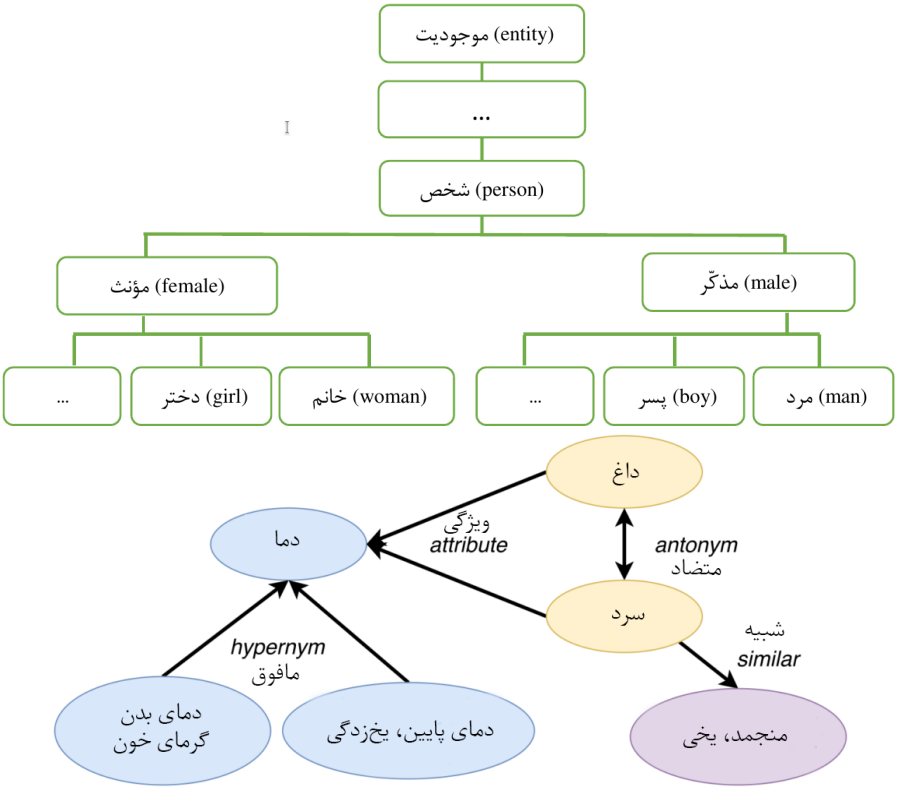
۱-۱۹ ساختار رساله

فصل دوم به بررسی روش‌های پیشین می‌پردازیم و مزیت‌ها و معایب هرکدام را بررسی کردیم.

فصل سوم به توضیح روش پیشنهادی می‌پردازیم و دلایل انتخاب روش‌ها را بررسی کردیم

فصل چهارم روش‌های پیشنهادی را آنالیز می‌کنیم و با آزمایش‌های انجام شده کارآمدی آن‌ها را اثبات کردیم

فصل پنجم خلاصه کوتاهی از پژوهش، روش‌های انجام شده و نتایج پژوهش را آورده‌ایم.



تصویر ۱-۱۲: نمونه‌ای از ساختار درختی شبکه واژگان [۳۲، ۸۳]

فصل ۲: کارهای پیشین

در این فصل به‌طورکلی، روش‌های محاسبه شباهت زوج‌جمله به دو قسمت تقسیم شده است. (۱) تلفیق روش‌های آماری و معیارهای مبتنی بر شبکه واژگان و (۲) مدل‌های عصبی. روش‌های مبتنی بر شبکه واژگان بسیار قابل اعتماد و قدرتمند هستند اما استفاده از این روش‌ها باعث وابسته شدن روش پیشنهادی به زبان مبنای شبکه واژه‌ها (مثلاً انگلیسی) خواهد شد. از طرفی دیگر، روش‌های فقط آماری وابسته به زبان مشخص نیستند و می‌توانند به زبان‌های دیگر به‌آسانی تعمیم پیدا کنند، هرچند بسیار زمان‌گیر و پیچیده هستند. همچنین جمع‌آوری اطلاعات کافی برای محاسبه شباهت دو واژه مشکل و برای دو جمله حتی سخت‌تر است. این فصل پیش‌زمینه روش‌های ارزیابی شباهت را معرفی می‌کند. علاوه بر این، موضوعات مطرح شده، نحوه بهبود ارزیابی شباهت را از طریق اضافه کردن مفاهیم[1] بررسی می‌کند.

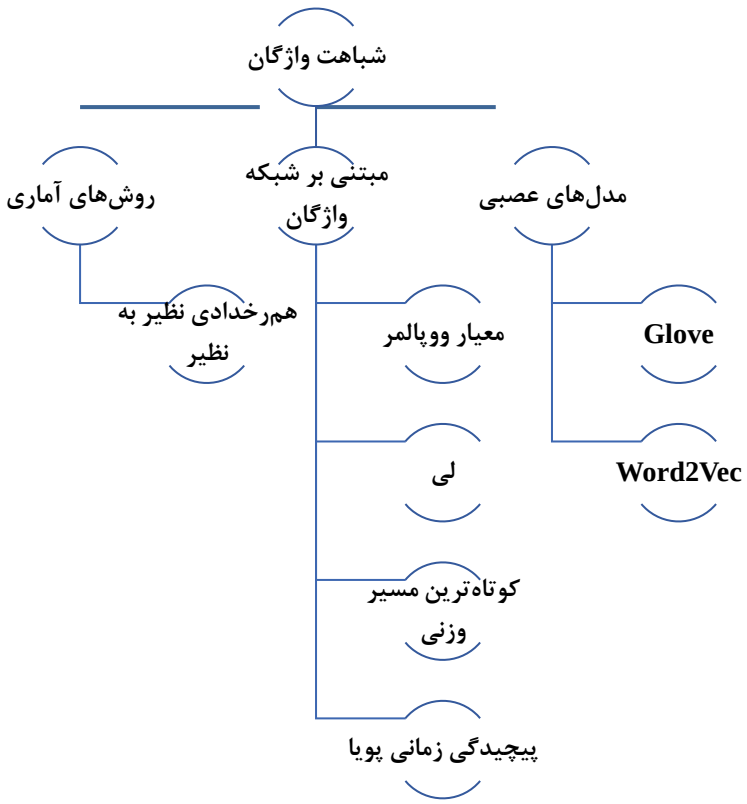
درک زبان انسان یکی از مهم‌ترین اهداف هوش مصنوعی از زمان ظهور آن بوده است. یکی از آزمون‌های کلاسیک برای ارزیابی هوشمندی، ماشین آزمون تورینگ است [۱۲۵] که در آن ماشین به محاوره با انسان می‌پردازد طوری که قابل تمایز با انسان نباشد. چندین حوزه پژوهشی مرتبط با این رساله، هم از نظر زبانی و هم از نظر محاسباتی معرفی می‌شوند. هرچند که مفاهیم کلیدی در علوم زبان، شناخته‌شده هستند اما بیشتر به معنای مستقل هر جمله می‌پردازند تا اینکه میزان شباهت یک جفت جمله ارزیابی شود. بنابراین، اینکه چطور این قواعد به‌منظور مقایسه جمله‌ها استفاده شوند، چارچوب پایه‌ای ساخت مدل‌های ارزیابی شباهت خواهند بود.

از کاربردهای گسترده ارزیابی شباهت متن می‌توان به خوشه‌بندی اسناد مشابه [۸۴]، خوشه‌بندی محتوای وب و جمله‌های کوتاه [۶۹، ۲۴]، متمایزسازی متون متفاوت [۳۴]، طرح پاسخ متناظر با پرسش کاربر در سیستم‌های پرسش و پاسخ خودکار [۳۶] و تشخیص سرقت ادبی [۱۲۷] اشاره کرد. همچنین اختلاف ترتیب واژگان بین دو جمله برای محاسبه مسائل استلزام، مهم شمرده شده است [۲]. در استلزام یک جمله به عنوان فرضیه داریم و سیستم باید تشخیص دهد که آیا جمله دوم در ارتباط با همین فرضیه است یا خیر [۱۲۶]. روش‌های محاسبه شباهت اسناد در پژوهشی توسط آلگِرِن و کُلی‌یاندِر، بررسی جامع شده است [۴]. در پژوهش ژانگ و چاو از ویژگی‌های استخراج شده در هر پاراگراف سند (محلی) و

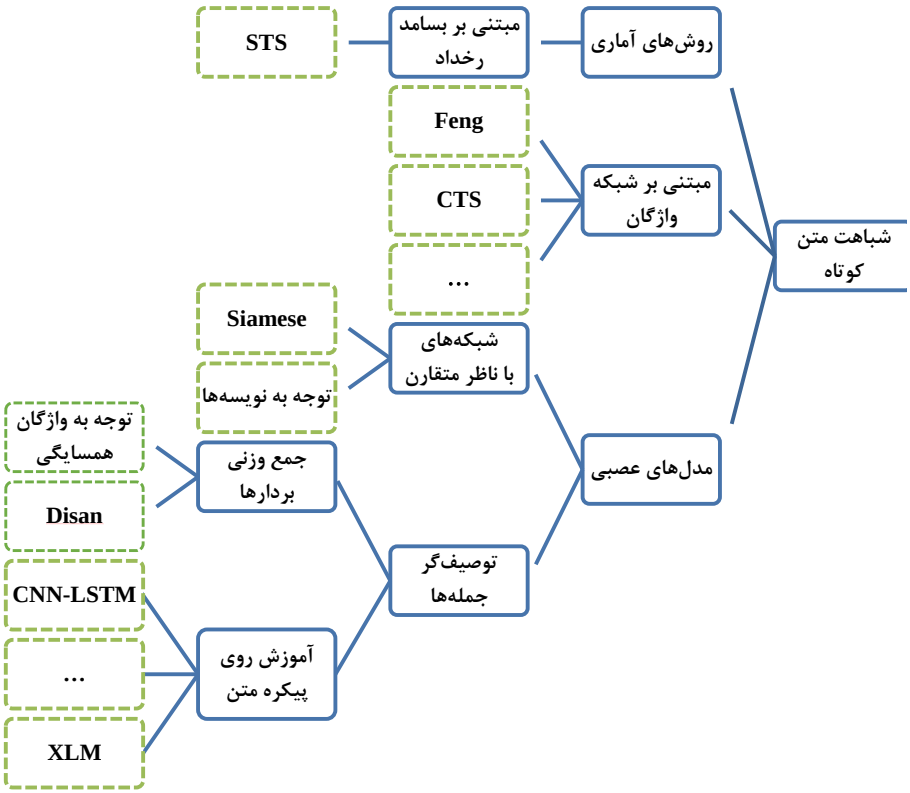
[1] Concepts

ویژگی‌های کل سند (عمومی) به‌منظور جستجوی اسناد مشابه استفاده شده است [۱۴۳]. در پژوهشی در سال ۲۰۱۹ از مترادف‌های موجود در سند که توسط شبکه واژگان استخراج می‌شود برای تشکیل بردار ویژگی و رده‌بندی اسناد استفاده شده است [۱۱۴]. در پژوهش لی و همکاران هر سند به عنوان گرافی از مجموعه مفهوم‌های موجود نمایش داده می‌شود، مفهوم‌ها با استفاده از بردار ویژگی آموزش دیده توسط شبکه عصبی، نمایش داده می‌شوند و ارتباط اسناد با استفاده از گراف‌ها محاسبه می‌شود [۹۳].

در ادامه در تصویر ۲-۱ و تصویر ۲-۲ روش‌های بررسی شده به صورت ساختار درختی نمایش داده شده است و پس از آن روش‌های آماری و سپس روش‌های پیشین مبتنی بر شبکه‌های عصبی به صورت جزئی بررسی شده‌اند.



تصویر ۲-۱: روش‌های پیشین بررسی شده از دیدگاه ارزیابی شباهت زوج‌واژگان



تصویر ۲-۲: روش‌های پیشین بررسی شده از دیدگاه ارزیابی شباهت زوج‌جمله‌ها

۲-۱ روش‌های آماری

۲-۱-۱ اطلاعات هم‌رخدادی نظیر به نظیر

روش اطلاعات متقابل نظیر به نظیر[۲] از اطلاعات هم‌رخدادی متون برای محاسبه احتمالات استفاده می‌کند. برای مثال در ساده‌ترین حالت آمار هم‌رخدادی دو واژه را در نظر می‌گیریم.

$$score(problem AND solutions) = log_2(p(problem\&solution)/p(problem)p(solution)) \quad (۲-۱)$$

در فرمول ۲-۱، هم‌رخدادی دو واژه محاسبه شده است. هرچقدر آمار صورت کسر بزرگ‌تر باشد، مقدار خروجی بیشتر می‌شود که نشان‌دهنده وابستگی آماری و هم‌رخدادی دو واژه است.

[2]Point-wise mutual information (PMI)

۲-۱-۲ آنالیز معنای نهفته

روش آنالیز معنای نهفته[۳] روشی پایه‌ای برای محاسبه شباهت زوج‌واژگان است. این روش مبتنی بر مقادیر بسامد رخداد واژگان موجود بین دو جمله است. به این معنی که ابتدا یک ماتریس محاسبه می‌شود که ستون‌های آن نمایانگر شناسه سند و ردیف‌های آن بسامد رخداد واژه در آن سند است. از ماتریس هم‌رخدادی توسط روش تجریه به مقدار ویژه [۴] اطلاعات ارزشمند استخراج می‌شود. به عبارتی، محور مختصات را به جهتی که نقاط بردار بسامد واژگان دارای بیشترین مقدار واریانس هستند تغییر می‌دهیم که نمونه مثال آن در تصویر ۲-۵ نمایش داده شده است. شباهت دو جمله در فضای تجزیه مقدار منفرد محاسبه می‌شود. این فضای مفهومی شباهت بسیاری با تعدادی از فن‌های ریاضی و آماری دارد که در زمینه‌های دیگری مورد استفاده قرار می‌گیرند، ازجمله: تجزیه بردارهای ویژه [۵] و آنالیز طیفی[۶]. فرض کنید در ابتدا واژگان سندها در مدل فضای برداری نمایش داده شود. این مدل در جدول ۲-۱ یکی از رویکردهای مبتنی بر مدل جبری برای نمایش اسناد متنی است که در صورت وجود یک واژه در سند عدد یک و در صورت نبود واژه عدد صفر تخصیص داده می‌شود. مشکل کلی این روش آن است که آنالیز معنای نهفته، یک مدل خطی است و نمی‌تواند روابط غیرخطی را مدل کند و علاوه بر آن، امکان تفسیر آسان نتایج وجود ندارد. سندهای نمونه زیر را در نظر بگیرید:

(a) An apple is a sweet, edible fruit produced by an apple tree.

(b) A peach is a soft, juicy and fleshy fruit.

(c) Grape is a soft fruit.

(d) A vehicle is a machine that transports people.

(e) A motorcycle is a vehicle.

[3] Latent semantic analysis (LSA)

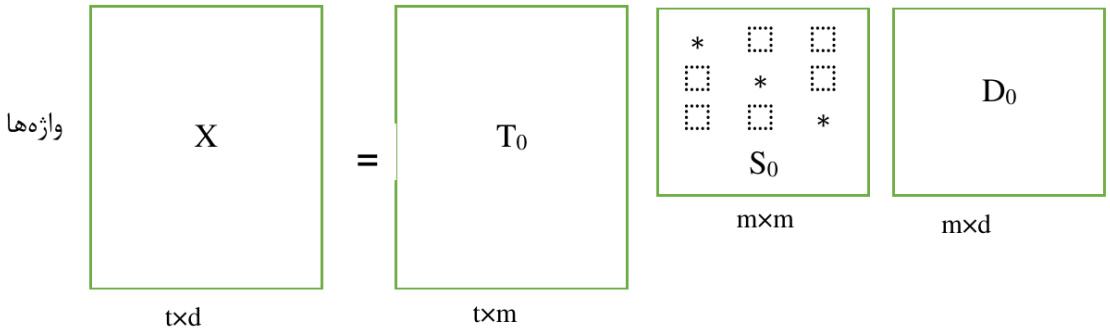
[4] SVD

[5] Eigenvector decomposition

[6] Spectral analysis

جدول ۲-۱: مدل فضای برداری

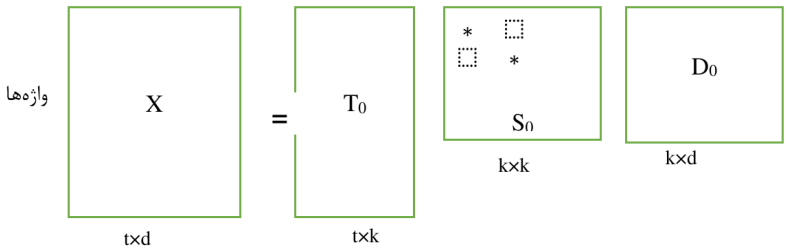
	a	b	c	d	e
apple	۱	۰	۰	۰	۰
sweet	۱	۰	۰	۰	۰
edible	۱	۰	۰	۰	۰
fruit	۱	۱	۱	۰	۰
produced	۱	۰	۰	۰	۰
tree	۱	۰	۰	۰	۰
peach	۰	۱	۰	۰	۰
soft	۰	۱	۱	۰	۰
juicy	۰	۱	۰	۰	۰
fleshy	۰	۱	۰	۰	۰
grape	۰	۰	۱	۰	۰
soft	۰	۰	۰	۰	۰
vehicle	۰	۰	۰	۱	۱
machine	۰	۰	۰	۱	۰
transport	۰	۰	۰	۱	۰
people	۰	۰	۰	۱	۰
motorcycle	۰	۰	۰	۰	۱



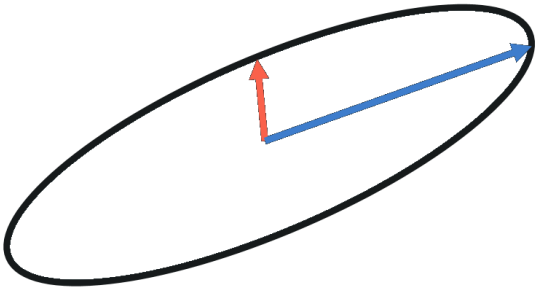
تصویر ۲-۳: تجزیه در فضای مقدار منفرد [۳۰]

بردار جدول ۲-۱ در فضای تجزیه مقدار منفرد به سه ماتریس مانند تصویر ۲-۳ تجزیه می‌شود که در آن:

- T_0 دارای ستون‌های متعامد با طول یکنواخت است به این معنا که $\acute{T}_0 T_0 = I$
- D_0 دارای ستون‌های متعامد با طول یکنواخت است به این معنا که $\acute{D}_0 D_0 = I$
- S_0 ماتریس قطری حاوی مقادیر ویژه است.
- t تعداد ردیف‌های X است.
- d تعداد ستون‌های X است.



تصویر ۲-۴: تشکیل فضای جدید با استفاده از فضای مقدار منفرد [۳۰]



تصویر ۲-۵: گردش محور مختصات به سمت بیشترین واریانس در داده‌ها [۱۵]

- m درجه[۷] x است و $m \leq min(t,d)$

اگر مقادیر ویژه در S_0 از بزرگ به کوچک مرتب باشند می‌توان فقط تعداد مورد نظر از بزرگ‌ترین مقادیر ویژه را نگهداشت و باقی را صفر گذاشت و بدین شکل به ماتریس تخمینی $\hat{x}$ با درجه $k \leq m$ دست پیدا کنیم (شکل ۲-۴).

۲-۱-۳ روش STS

روش STS شباهت معنایی دو متن را با استفاده از معیارهای آماری مبتنی بر بسامد رخداد واژگان محاسبه می‌کند و طولانی‌ترین دنباله‌ی نویسه‌ای مشترک (NLCS) بین زوج‌واژه را نیز در نظر می‌گیرد[۵۹]. ایده NLCS این است که غلط املایی نویسه یک واژه باید توسط روش پیشنهادی تشخیص داده شود. برای مثال در دو جمله:

- $T1$: Many consider Maradona as the best player in soccer history.
- $T2$: Maradena is one of the best soccer players.

[7] Rank

واژه Maradona در جمله دوم اشتباه نوشته شده است اما این دو واژه‌یکسان در نظر گرفته شوند. ایده بعدی روش STS این است که واژگان مشابه هم، در سندهای مشابه هم‌رخداد هستند. به عبارتی دیگر واژه «گربه» و «سگ» به همدیگر شبیه هستند زیرا در اطراف هردوی این واژگان به احتمال زیاد واژه‌هایی مشترک مانند «حیوانات»، «غذا» وجود دارد. از طرفی دیگر، واژه «گربه» به «میز» شبیه نیست زیرا این دو واژه معمولاً در سندهای متفاوتی نوشته می‌شوند. ایده سوم این است که ترتیب واژگان در جمله‌های شبیه به هم، مانند یکدیگر است. برای مثال، دو جمله‌ی قبل را (بدون اشتباه املایی) در نظر بگیرید؛ بعد از حذف ایست‌واژه‌ها و ریشه‌یابی، پیدا کردن جزء اصلی واژه، شکل واژه بدون پیشوند یا پسوند به دو لیست زیر می‌رسیم:

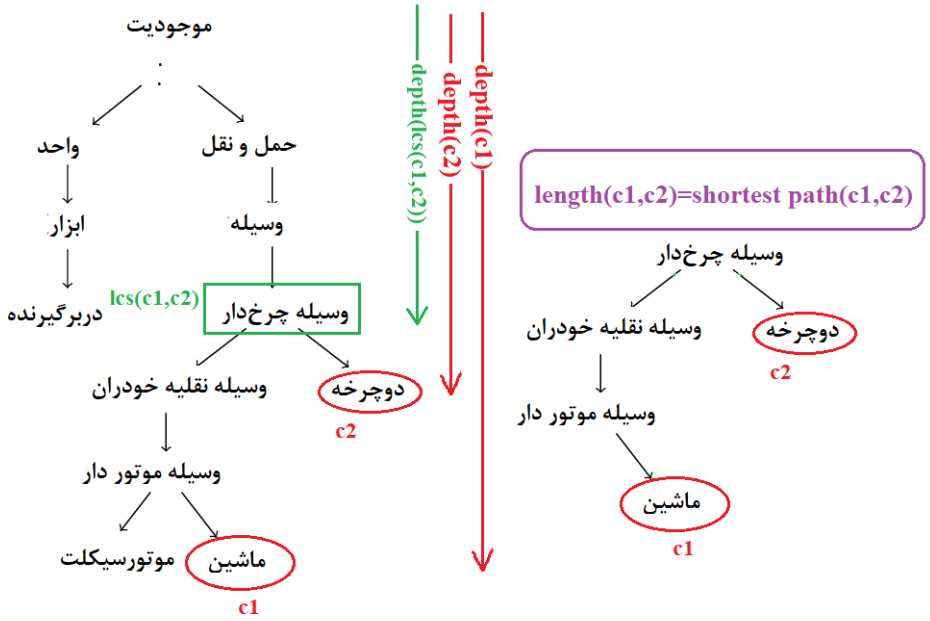
- $T1$ = maradona, the, best, player, soccer
- $T2$ = maradona, the, best, soccer, player

برای محاسبه شباهت ترتیب واژگان زوج‌جمله، ابتدا واژگان مشترک بین دو جمله شناسایی می‌شود، سپس به تمام واژگان جمله‌ی اول عددی به عنوان عدد شاخص تخصیص داده می‌شود و با توجه به آن برای جمله دوم عددهای شاخص تخصیص داده می‌شود بنابراین داریم:

- $x = \{1, 2, 3, 4, 5\}$
- $y = \{1, 2, 3, 5, 4\}$

در نتیجه، نحوه محاسبه شباهت ترتیب واژگان بین دو جمله طبق فرمول ۲-۲ است که در آن n تعداد واژه‌های مشترک بین دو جمله است.

$$sim_{order} = \left\{ \begin{array}{cc} 1 - \frac{2\sum_{i=1}^{n} |x_i - y_i|}{n^2} & if\ n \rightarrow \text{زوج} \\ 1 - \frac{2\sum_{i=1}^{n} |x_i - y_i|}{n^2 - 1} & if\ n \rightarrow \text{فرد}\ and\ n > 1 \\ 1 & if\ n \rightarrow \text{فرد}\ and\ n = 1 \end{array} \right\} \quad (۲-۲)$$



تصویر ۲-۶: نمونه‌ای از نحوه محاسبه فاصله‌ها در شبکه واژگان [۱۰۰]

۲-۲ شباهت مبتنی بر ساختار درختی شبکه واژگان

هدف از پیشنهاد این معیارهای شباهت، ابهام‌زدایی از معنای واژه است. برای مثال، شباهت دو واژه «market» و «apple» را در نظر بگیرید. معیارهای مبتنی بر شبکه واژگان، نزدیکترین مفهوم مشترک بین این دو را پیدا می‌کند، در نتیجه برای محاسبه شباهت از واژه «apple» ابهام‌زدایی می‌شود و معنای «شرکت اپل» مورد استفاده قرار می‌گیرد (در مقابل میوه سیب). در تصویر ۲-۶ نمونه‌ای از شبکه واژگان را مشاهده می‌کنید و تعدادی از معیارهای فاصله در آن مشخص شده است.

۲-۲-۱ معیار وو و پالمر

این معیار شباهت، موقعیت مکانی مفهوم‌های c_1 و c_2 و نزدیک‌ترین مفهوم مشترک این دو را در نظر می‌گیرد [۱۳۷].

$$sim_{wup}(c_1, c_2) = \frac{2 \times depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)} \quad (۳-۲)$$

محاسبه شباهت با عمق نزدیکترین مفهوم مشترک، رابطه مستقیم دارد و توسط جمع فاصله هر مفهوم از ریشه شبکه واژگان، به عددی بین صفر و یک نگاشت می‌شود. مزیت این روش، سادگی محاسباتی آن است، اما در آن هیچ‌گونه اطلاعات آماری استفاده نمی‌شود و فاصله کوتاه‌ترین مسیر بین دو مفهوم بررسی نشده است.

۲-۲-۲ معیار لی

برای مثال، واژه «depression» اگر در بافت متن اقتصادی به کار رود باید به معنی رکود در نظر گرفته شود و اینکه بازار به مدت طولانی در یک روند نزولی قرار گرفته است. به نظر می‌رسد در محاسبه شباهت، می‌توان با معیار کوتاه‌ترین مسیر بین زوج‌واژه، در ابهام‌زدایی از معنای واژه عملکرد بهتری نسبت به روش ووپالمر داشت؛ زیرا که معیار لی، فاصله مستقیم بین دو واژه را نیز در نظر می‌گیرد. این معیار، به‌طور شهودی و تجربی مشتق شده است و تابعی از کوتاه‌ترین مسیر دو مفهوم c_1 و c_2 با عمق نزدیک‌ترین مفهوم مشترک این دو، نسبت به ریشه شبکه واژگان است [۷۵]. در فرمول ۲-۴، $\alpha \geq 0$ و $\beta > 0$ پارامترهایی هستند که به ترتیب، نسبت تأثیر کوتاه‌ترین مسیر و عمق را تنظیم می‌کنند. طبق پژوهش لی و همکاران [۷۵]، بهترین پارامترها $\alpha = 0/2, \beta = 0/6$ هستند.

$$sim_{li}(c_1, c_2) = e^{-\alpha \times length} \frac{e^{\beta \times depth(LCS(c_1,c_2))} - e^{-\beta \times depth(LCS(c_1,c_2))}}{e^{\beta \times depth(LCS(c_1,c_2))} + e^{-\beta \times depth(LCS(c_1,c_2))}} \quad (۴-۲)$$

دو روش بررسی شده معیارهای بالا، تنها بر مبنای پیوندهای $(IS-A)$ بین مفاهیم قرار دارند، با

فرض این که پیوندها در رده‌بندی، نشان‌دهنده فواصل هستند، اما چگالی عبارات در رده‌بندی عموماً ثابت نیست. به‌طور معمول، عبارات کلی‌تر که در سلسله‌مراتب بالاتر وجود دارند، با مجموعه کوچک‌تری از گره‌ها مرتبط هستند؛ در حالی که تعداد بیشتری از واژگان خاص‌تر با چگالی بسیار بالاتر در سلسله‌مراتب پایین‌تری قرار دارند. برای مثال فاصله لینک بین «گیاه» و «حیوان» دو است (والد مشترک آن‌ها موجود زنده است). و فاصله بین «گورخر» و اسب نیز دو است (والد مشترک آن‌ها اسب‌مانند است). به‌طور شهودی اسب و گورخر به هم بیشتر مرتبط هستند تا گیاه و حیوان، این در حالی است که این دو فاصله، یکسان در نظر گرفته می‌شود.

۲-۲-۳ معیار کوتاه‌ترین مسیر وزنی

بیشتر روش‌های مبتنی بر شبکه واژگان، یا فقط روی محاسبه محتوای اطلاعات واژگان برای ارزیابی شباهت تمرکز کرده‌اند و یا با استفاده از محاسبه کوتاه‌ترین مسیر بین دو واژه، شباهت را تخمین می‌زنند. در این معیار، هردوی ویژگی‌های نامبرده برای محاسبه شباهت، تلفیق می‌شوند تا معیار دقیق‌تری طراحی شود. این معیار با استفاده از فرمول ۲-۵ محاسبه می‌شود.

$$sim_{wpath}(c_1, c_2) = \frac{1}{1 + length(c_1, c_2) \times \alpha^{IC(C_{lcs})}} \qquad (۵-۲)$$

در این فرمول، محاسبه میزان تأثیر اطلاعات بسامدی دو واژه، با استفاده از متغیر α تنظیم می‌شود. C_{LCS} نزدیک‌ترین مفهوم مشترک بین دو مفهوم مورد مقایسه است. در این فرمول، $length(c_1, c_2)$ فاصله کوتاه‌ترین مسیر بین دو واژه است [۱۴۵].

۲-۳ محاسبه شباهت دو جمله مبتنی بر شبکه واژگان

۲-۳-۱ Feng

در روش فِنگ ژو و مارتین، دو روش برای محاسبه شباهت دو جمله استفاده شده است: ارتباط مستقیم و غیرمستقیم. در نوع اول، ابتدا هر جمله با استفاده از واژه‌های مترادف بسط داده می‌شود [۳۵]. به عبارتی، اگر که جمله مانند فرمول ۲-۶ باشد، جمله‌ی بسط داده شده مانند فرمول ۲-۷ خواهد بود:

$$\begin{aligned} s_1 &= \{NN_1, VB_1, ADJ_1, ADV_1\} \\ s_2 &= \{NN_2, VB_2, ADJ_2, ADV_2\} \end{aligned} \tag{۲-۶}$$

$$\begin{aligned} \acute{s}_1 &= \{\acute{N}N_1, \acute{V}B_1, \acute{A}DJ_1, \acute{A}DV_1\} \\ \acute{s}_2 &= \{\acute{N}N_2, \acute{V}B_2, \acute{A}DJ_2, \acute{A}DV_2\} \end{aligned} \tag{۲-۷}$$

شباهت دو واژه با فرمول ۲-۳ محاسبه شده است، سپس مانند فرمول ۲-۸، شباهت هر واژه در جمله اول با واژه‌های هم‌نقش در جمله دوم محاسبه می‌شود. در مرحله بعد، اهمیت هر «نقش‌واژه» در جمله اول، نسبت به جمله دوم محاسبه می‌شود و در ادامه، با ضرب‌در محتوای اطلاعات[8]، واژه طبق فرمول ۲-۹ وزن‌دهی می‌شود.

$$sim(w_i|s_2) = \left\{ \begin{array}{ll} \sum_{q_k \in NN_2} \frac{sim(w_i, q_k)}{wordPairNumber} & if|\acute{N}N_1| > 0, |\acute{N}N_2| > 0 \\ 0.05 & if|\acute{N}N_1| > 0, |\acute{N}N_2| = 0 \end{array} \right\} \tag{۲-۸}$$

[8] Information content

در فرمول ۲-۸ نماد q_k به معنای واژه شماره k در دسته واژگان متعلق به گروه اسم در جمله دوم NN_2 است. به عبارتی دیگر شباهت واژه w_i در جمله اول با واژه‌های گروه اسم جمله دوم محاسبه می‌شود. سپس، معادله توسط wordPairNumber که نماد تعداد کل زوج‌واژگان بین دو جمله است نرمال می‌شود. در ادامه، طبق فرمول ۲-۹، به ازای تمام واژه‌های w_i مقادیر جمع می‌شوند.

$$sim(\acute{N}N_1|s_2) = \sum_{W_1 \in \acute{N}N_1} \frac{1}{|\acute{N}N_1|} sim(w_i|s_2) IC(w_i)$$
$$IC(w) = 1 - \frac{log(n+1)}{log(N+1)} \quad (۹-۲)$$

که در آن n بسامد واژه w در پیکره و N جمع بسامد تمام واژه‌ها می‌باشد و نتیجه نهایی برابر با جمع شباهت‌های زوج‌واژه‌ها است. شباهت شرطی جمله اول به دوم، به صورت زیر محاسبه می‌شود.

$$sim(s_1|s_2) = \alpha(sim(\acute{N}N_1|s_2) + sim(\acute{V}B_1|s_2)) + \beta(sim(\acute{A}DJ_1|s_2) + sim(\acute{A}DV_1|s_2)) \quad (۱۰-۲)$$

که در آن $\alpha = 0/1$ و $\beta = 0/8$ اهمیت هر دسته نقش‌واژه را در محاسبه مشخص می‌کند. در انتها، شباهت دو جمله از طریق «ارتباط مستقیم» با فرمول زیر محاسبه می‌شود.

$$sim_{direct}(s_1, s_2) = \frac{|\acute{s}_1|}{|\acute{s}_1| + |\acute{s}_2|} sim(s_1|s_2) + \frac{|\acute{s}_1|}{|\acute{s}_1| + |\acute{s}_2|} sim(s_2|s_1) \quad (۱۱-۲)$$

که در آن $|\acute{s}_1|$ تعداد واژه‌های جمله بسط داده شده اول و $|\acute{s}_2|$ طول جمله دوم است. طبق این پژوهش، شباهت بین دو جمله فقط به واژه‌های مورد استفاده بستگی ندارد بلکه وابسته به شرایط، دنباله‌ی مشخصی از واژه‌ها مورد استفاده قرار می‌گیرد.

محاسبه میزان هم‌ترازی‌هم‌ترازی دو دنباله، ارتباط غیرمستقیم بین دو جمله را تشکیل می‌دهد که

با استفاده از الگوریتم نیدِلمَن-وانچ[9] محاسبه می‌شود [۹۱]. به عبارتی چه تعداد ویرایش لازم است تا یک جمله به جمله‌ای دیگر تبدیل شود. این الگوریتم، در ابتدا طراحی شد تا بتوان دو توالی پروتئین یا نوکلئوتیدی باهم هم‌تراز [10] شوند. در فرمول ۲-۱۲ نماد c_{1i} به عنوان واژه i در جمله اول است. در این فرمول، برای دو رشته ورودی c_1 و c_2 با طول به ترتیب M و N، ماتریس F تشکیل می‌گردد که درایه $F_{i,j}$ آن تراز بهینه $c_{1i...1M}$ و $c_{2j...2N}$ است. این الگوریتم از بخش بالای راست ماتریس تا درایه پایین چپ پر می‌شود. نحوه محاسبه هر درایه ماتریس در فرمول زیر آورده شده است. d پارامتر تنبیه روش، به دلیل عدم هماهنگی بین دو دنباله رشته است و مقدار آن قابل تنظیم است.

$$F_{i,j} = max \begin{Bmatrix} F_{i-1,j-1} + sim_{wup}(c_{1i}, c_{2j}) \\ F_{i-1,j} - d \\ F_{i,j-1} - d \end{Bmatrix} F_{0,0} = 0 \qquad (۱۲-۲)$$

که در آن $s(c_{1i}, c_{2j})$ شباهت بین دو واژه است که مطابق با معیار ووپالمر فرمول ۲-۳ محاسبه می‌شود، سپس به شکل فرمول ۲-۱۳ وزن‌دهی می‌شود.

$$sim_{wup}(c_{1i}, c_{2j}) = max \begin{Bmatrix} \theta sim_{wup}(c_{1i}, c_{2j}) & sim_{wup}(c_{1i}, c_{2j}) \geq \zeta \\ -1 & sim_{wup}(c_{1i}, c_{2j}) < \zeta \end{Bmatrix} \qquad (۱۳-۲)$$

مقادیر ζ و θ پارامترهای قابل تنظیم هستند که مقدار θ وابسته به نقش دو واژه مورد مقایسه است. مقدار نهایی «ارتباط غیرمستقیم» دو جمله در فرمول زیر آورده شده است.

[9] Needleman-wunsch

[10] Align

$$sim_{indirect}(s_1, s_2) = \frac{F_{|s_1|,|s_2|}}{max(|s_1|, |s_2|)} \quad (۱۴-۲)$$

که در آن s_1 و s_2 تعداد واژه‌های جمله اول و دوم است. فرمول نهایی محاسبه شباهت دو جمله مطابق با فرمول زیر است.

$$\begin{aligned} sim_{feng}(s_1, s_2) = function(dr, indr) = \lambda sim_{direct} + (1-\lambda) sim_{indirect} \\ = \lambda(\frac{|\acute{s}_1|}{|\acute{s}_1|+|\acute{s}_2|} sim(s_1|s_2) + \frac{|\acute{s}_1|}{|\acute{s}_1|+|\acute{s}_2|} sim(s_2|s_1) + (1-\lambda)\frac{F_{|s_1|,|s_2|}}{max(|s_1|,|s_2|)} \end{aligned} \quad (۱۵-۲)$$

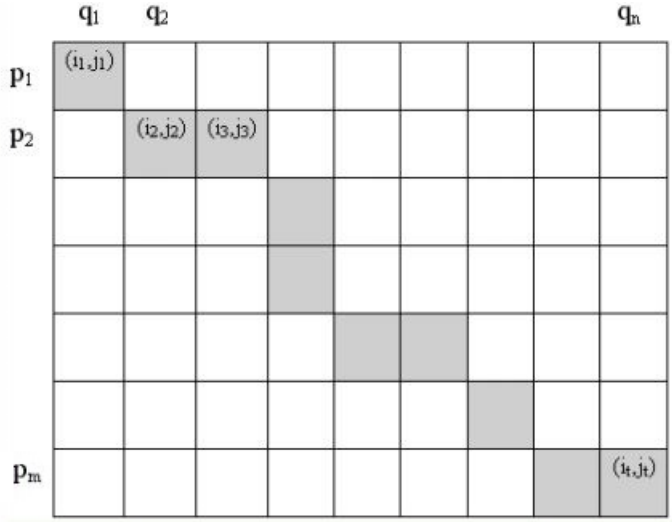
پارامتر λ تنظیم‌کننده‌ی میزان تأثیر «ارتباط مستقیم» و «غیرمستقیم» بین دو جمله است.

۲-۳-۲ پیچیدگی زمانی پویا

در پژوهشی توسط لیو، ژو و ژِنگ، پژوهشگران از الگوریتم پیچیدگی زمانی پویا[11] (با کارکرد مشابه نیدلمن-وانچ) برای محاسبه شباهت متن استفاده کرده‌اند [۷۷]. این الگوریتم برای پیدا کردن دو سیگنال مشابه که یکی در طول زمان، سریع‌تر از دیگری است استفاده می‌شود. به‌طورکلی می‌تواند برای پیدا کردن شکل‌های یکسان که حاوی تغییرات کمی هستند استفاده شود. فرض کنید دو رشته $p =< p_1, p_2, ..., p_m >$ و $q =< q_1, q_2, ..., q_n >$ حاوی m و n واژه باشند؛ یک ماتریس $m \times n$ شکل می‌گیرد و مسیر w در فرمول ۲-۱۶ دنباله‌ای از درایه‌های ماتریس است که تراز بین P و Q را تعیین می‌کند.

$$\begin{aligned} w = (i_1, j_1), (i_2, j_2), ..., (i_t, k_t) \\ max(m, n) \leq t < m + n - 1 \end{aligned} \quad (۱۶-۲)$$

[11] Dynamic-programming-language



تصویر ۲-۷: مسیر هم‌ترازی دو جمله با الگوریتم زمانی پویا [۱۳]

تصویر ۲-۷ این مسیر را در ماتریس نمایش می‌دهد.

مقادیر هر درایه در ماتریس با فرمول ۲-۱۷ محاسبه می‌شود.

$$f(i,j) = dist(p_i, q_j) + min\left\{\begin{array}{c} f(i-1,j) \\ f(i,j-1) \\ f(i-1,j-1) \end{array}\right\} \qquad (۱۷-۲)$$

$$f(0,0) = 0$$

$$f(i,0) = f(0,j) = \infty$$

$$i \in (0,m), j \in (0,n)$$

تابع dist شباهت بین دو واژه را با فرمول ۲-۱۸ محاسبه می‌کند.

$$dist(w_1, w_2) = 1 - \frac{e^{\alpha \times depth(LCS(c_1,c_2))} - 1}{e^{\alpha \times depth(LCS(c_1,c_2))} + e^{\beta \times length(c_1,c_2)} - 2} \qquad (۱۸-۲)$$

که در آن α و β پارامترهای قابل تنظیم هستند [۷۷] و مقادیر پیشنهادی آن‌ها توسط پژوهش لیو و

همکاران $\alpha = 0/1$ و $\beta = 0/45$ پیشنهاد شده است. برای یافتن مسیر بهینه از (i_1, j_1) تا (i_t, j_t) می‌توان از روش حریصانه یا الگوریتم برنامه‌نویسی پویا استفاده کرد [۱۳]. مسیر بهینه مسیری است که جمع تجمعی درایه‌های موجود در آن کمینه باشد. سپس از این مقدار کمینه، به عنوان تعداد ویرایش مورد نیاز برای تبدیل یک جمله به جمله دیگر در نظر گرفته می‌شود.

۲-۳-۳ LSS و HLA

روش HAL[12] نیز مشتق شده از آنالیز معنای نهفته است؛ البته نتایج تجربی نشان می‌دهد که این روش برای محاسبه شباهت معنایی متون کوتاه مناسب نیست [۲۲]. برای روش LSS شباهت دو واژه با پیدا کردن «مقدار شباهت بیشینه» در «مسیر متصل‌کننده این دو در شبکه واژه» محاسبه می‌شود؛ به عبارتی کوتاه‌ترین مسیری که دو واژه را به هم متصل کند. دو جمله زیر را در نظر بگیرید [۲۹]:

- The chrysanthemum lady
- A woman selling flowers

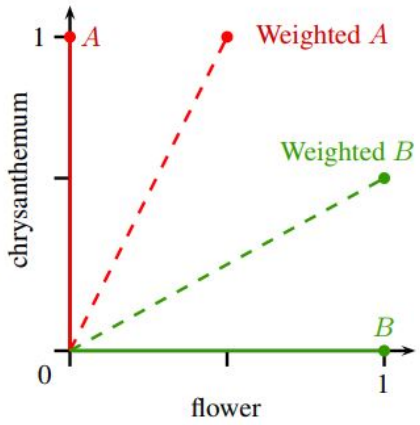
معنای این دو جمله به ترتیب «زنی که گل داوودی می‌فروشد» و «خانمی که گل می‌فروشد» است.این دو جمله بعد حذف ایست‌واژه‌ها و جدا کردن واژه‌ها حاوی دو لیست زیر هستند:

- chrysanthemum, lady
- woman, selling, flowers

اگر شباهت واژه‌ها را باهم در نظر بگیریم به تصویر ۲-۸ می‌رسیم. در نتیجه به این شکل می‌توان با مشکل کمبود داده برای محاسبه شباهت راه‌حلی ارائه داد. اگر که حالا فاصله کسینوس دو واژه گل (flower) و گل داوودی (chrysanthemum) را محاسبه کنیم مقدار غیر صفر خواهیم داشت. میزان شباهت بین تمامی زوج واژه‌ها مانند جدول ۲-۲ محاسبه می‌شود

در پی آن، بردار هریک از جمله‌های اول و دوم به صورت جدول ۲-۳ محاسبه می‌شود.

[12] Hyperspace analogs to language



تصویر ۲-۸: وزن‌دهی به واژه‌ها برای غیر صفر کردن میزان شباهت آن‌ها [۲۹]

جدول ۲-۲: ماتریس شباهت واژگان

woman	selling	lady	flower	chrysanthemum	
۰/۱	۰/۰۶	۰/۰۹	۰/۵	۱	chrysanthemum
۰/۱۱	۰/۰۹	۰/۱	۱		flower
۰/۵	۰/۰۵	۱			lady
۰/۰۹	۱				selling
۱					woman

جدول ۲-۳: شکل‌گیری بردار جمله‌ها

woman	selling	lady	flower	chrysanthemum		
۰	۰	۱	۰	۱	جمله اول	بردارهای واژگان
۱	۱	۰	۱	۰	جمله دوم	
۰/۱	۰/۰۶	۰/۰۹	۰/۵	۱	chrysanthemum	جمله اول
۰/۵	۰/۰۷	۱	۰/۱	۰/۰۹	lady	
۰/۱۱	۰/۰۹	۰/۱	۱	۰/۵	flower	جمله دوم
۰/۰۹	۱	۰/۰۷	۰/۰۹	۰/۰۶	selling	
۱	۰/۰۹	۰/۵	۰/۱۱	۰/۱	woman	
۰/۶	۰/۱۳	۱/۰۹	۰/۶	۱/۰۹	جمله اول	بردار وزنی
۱/۲	۱/۱۸	۰/۶۷	۱/۲۰	۰/۶۶	جمله دوم	

جدول ۲-۴: محاسبه بردار ویژگی جمله در روش STASIS

	RAM	keeps	things	being	worked	with	the	CPU	uses	as	a	short-term	memory	store
RAM	1												0/8147	0/8147
keeps		1												
things			1					0/2802	0/4433					
being				1										
worked					1									
with						1								
vec_1	1	1	1	1	1	1	0	0/2843	0/4433	0	0	0	0/8147	0/8147

همان‌طور که در جدول ۲-۳ آورده شده است، جمع شباهت تمام واژه‌های جمله اول با واژه اول جمله بعدی به عنوان یک ویژگی محاسبه می‌شود. ویژگی‌های دیگر نیز به همین صوزت استخراج می‌شوند که در جدول برجسته شده است. مقدار تفاوت دو جمله فاصله، کسینوس بردار وزنی متناظر آن‌ها می‌باشد.

STASIS ۴-۳-۲

روش STASIS معرفی شده توسط لی و همکاران از دو بخش محاسبه شباهت معنایی و مقایسه ترتیب واژگان تشکیل شده است [۷۶].

۱) شباهت معنایی

ابتدا دو جمله مورد مقایسه باهم در یک جمله قرار می‌گیرند، برای مثال دو جمله:

- s_1: RAM keeps things being worked with.
- s_2: The CPU uses RAM as a short-term memory store.

با اتصال به هم تبدیل به جمله زیر می‌شوند:

T: RAM keeps things being worked with The CPU uses as a short-term memory store

جدول محاسبه بردار ویژگی جمله اول در جدول ۲-۴ نمایش داده شده است. همان‌طور که مشخص شده است، برای هر واژه «در ردیف اول جمله T»، واژه‌ای ازجمله s_1 که بیشترین شباهت را دارد پیدا می‌شود؛ سپس مقدار شباهت آن به عنوان یک ویژگی در بردار جمله که vec_1 است ذخیره می‌گردد. بردار جمله دوم نیز به همین ترتیب محاسبه می‌شود سپس معیار کاسینوسی این دو بردار به عنوان معیار شباهت دو جمله مورد استفاده قرار می‌گیرد.

۲) ترتیب واژگان

دو جمله زیر را در نظر بگیرید:

- s_1: A quick brown dog jumps over the lazy fox.
- s_2: A quick brown fox jumps over the lazy dog.

ابتدا یک جمله تلفیقی از این دو جمله که حاوی واژگان تکراری نباشد به صورت زیر تشکیل می‌شود:

T = a quick brown dog jumps over the lazy fox

سپس برای هر واژه در جمله s_1 و s_2، با توجه به مکان قرارگیری آن در جمله ترکیبی، عددی تخصیص داده می‌شود که در نتیجه آن بردارهای زیر به دست می‌آید:

$$r_1 = 1, 2, 3, 4, 5, 6, 7, 8, 9$$

$$r_2 = 1, 2, 3, 9, 5, 6, 7, 8, 4$$

سپس، با استفاده از فرمول ۲-۱۹ شباهت ترتیب واژگان بین دو جمله حساب می‌شود.

$$s_r = 1 - \frac{||r_1 - r_2||}{||r_1 + r_2||} \qquad (۱۹-۲)$$

روش نهایی محاسبه شباهت مطابق با فرمول ۲-۲۰ است که در آن ضریب δ در شباهت کاسینوس بردار دو جمله با شباهت ترتیب واژگان جمع می‌شود.

$$sim_{stasis}(s_1, s_2) = \delta(\frac{vec_1 \cdot vec_2}{||vec_1|| \cdot ||vec_2||}) + (1 - \delta)\frac{||r_1 - r_2||}{||r_1 + r_2||} \qquad (۲۰-۲)$$

Omiotis ۵-۳-۲

روش معرفی شده توسط سات‌سِرانیس و وِزیریانیس، ارزیابی دیگری از وابستگی بین متون فراهم می‌کند و معیار ارتباط معنایی واژه به واژه را برای ارزیابی بین متون تعمیم می‌دهد [۱۲۴]. در روش Omiotis از شبکه واژگان به عنوان منبع پیوندهای معنایی استفاده شده است. هر جفت از واژگان به‌طور بالقوه از طریق یک یا چند مسیر معنایی به هم متصل می‌شوند. Omiotis ارتباط دو واژه را با اطلاعات آماری واژگان ادغام می‌کند تا امتیاز همبستگی معنایی نهایی بین متون را فراهم کند. بنابراین دستاوردهای آن را می‌توان به صورت زیر خلاصه کرد:

۱. همه نوع روابطی که ساختار شبکه واژگان می‌تواند فراهم کند از جمله نقش واژگان، همچنین وزن‌دهی تأثیر این روابط و عمق مفهوم مشترک برای محاسبه شباهت دو واژه استفاده شده است که شانس پیدا کردن مسیر معنایی بین دو واژه را افزایش می‌دهد.

۲. یک معیار جدید برای محاسبه ارتباط معنایی بین متن معرفی شده است که نیازی به استفاده از پیکره‌های خارجی یا روش‌های یادگیری باناظر و یا بدون ناظر ندارد.

۳. در دسترس بودن امتیازات وابستگی معنایی که از قبل بین هر جفت از واژگان ارزیابی شده است، محاسبه ارتباط معنایی بین متن را سرعت می‌بخشد.

CTS ۶-۳-۲

روش CTS [۱۳] بررسی می‌کند که چون می‌توان اجزای مختلف جمله: «فعل، صفت یا قید» را به نقش‌واژه نام متناظر آن‌ها در شبکه واژگان نگاشت کرد، بدین ترتیب می‌توان مفهوم مشترک بین دو واژه را نیز پیدا کرد[۵۴]. همچنین اشاره می‌کند که اگر واژه‌ای به‌وفور در سندها استفاده شود، بار اطلاعاتی کمتری نسبت به واژه‌هایی دارد که کمتر در سندها دیده می‌شوند. این ارزش با استفاده از «معیار محتوای اطلاعاتی» محاسبه می‌شود و مستقیماً در ارزیابی شباهت جفت واژه استفاده شده است. برای مثال

[13] Concept transferred space

واژه «مرد» از «شخص» ارزش اطلاعاتی بیشتری دارد چراکه «شخص» به عنوان والد «مرد» در شبکه واژه شناخته می‌شود. نام موجودیت‌ها توسط ابزار آماده استنفرد شناسایی می‌شوند. دستاوردهای این پژوهش را می‌توان به صورت زیر خلاصه کرد:

۱. روش جدیدی برای نگاشت واژه‌های غیر اسم معرفی شده است طوری که تمامی نقش‌واژه‌ها بتوانند در محاسبه شباهت استفاده شوند.

۲. روش جدیدی برای کاهش از دست رفتن اطلاعات، با استفاده از نام موجودیت‌ها در محاسبه شباهت، معرفی شده است.

۳. روشی برای تعیین تأثیر واژه‌های مختلف در ارزیابی شباهت معنایی، در نظر گرفته شده است.

در این روش، مفهوم مشترک بین دو واژه از فرمول ۲-۲۱ و محتوای اطلاعات دو واژه مطابق فرمول ۲-۲۲ محاسبه می‌شود.

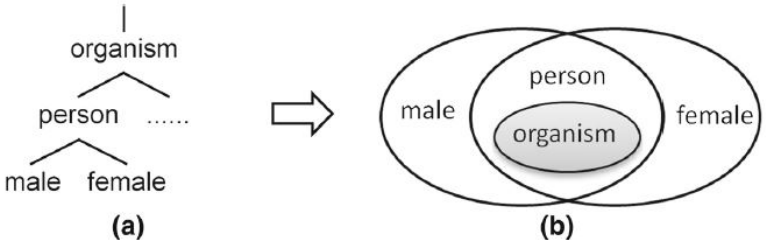
$$commonIC(c_1, ..., c_n) = IC[\cap_{i=1}^{n} c_i] \qquad (۲۱-۲)$$

$$IC(c_1, c_2) = IC(c_1) + IC(c_2) - commonIC(c_1, c_2) \qquad (۲۲-۲)$$

در تصویر ۲-۹ نمونه‌ای از روابط ساختاری بین مفهوم‌ها در شبکه واژگان آورده شده است. همان‌طور که در تصویر مشاهده می‌کنید، برای مثال مفهوم مشترک بین male و female ← person است و مفهوم مشترک بین male و person ← organism است. نحوه محاسبه هر دو به ترتیب به صورت زیر می‌باشد:

IC(male,female)=IC(male)+IC(female)-commonIC(male,female)

commonIC(male,female)=IC(person)



تصویر ۲-۹: مثال روابط ساختاری بین مفهوم‌های مختلف [۵۴]

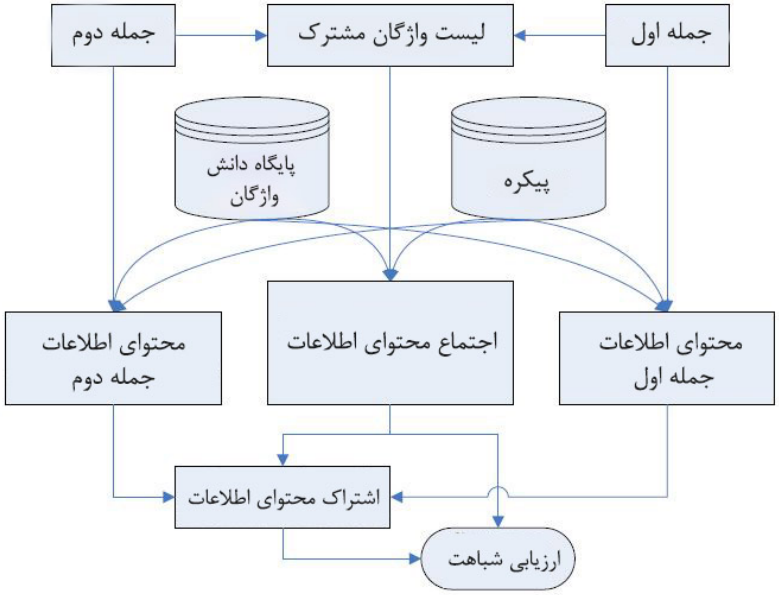
IC(male,person)=IC(male)+IC(person)-commonIC(male,person)

commonIC(male,person)=IC(organism)

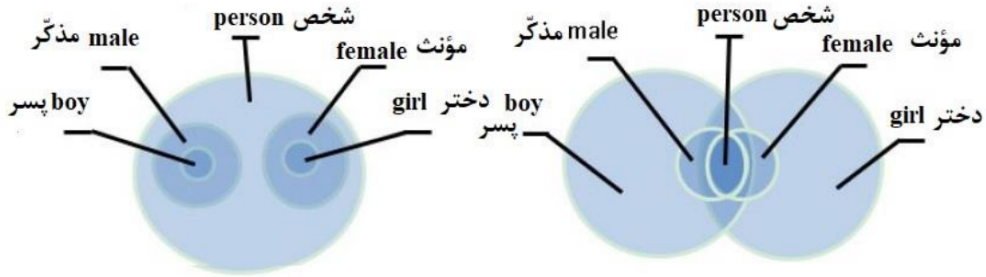
روش CTS شباهت را با در نظر گرفتن «همپوشانی محتوای اطلاعات» بین «دو جمله مورد مقایسه» محاسبه می‌کند. جمله، به عنوان منبع محتوای اطلاعات است و هر واژه در آن شامل معنی‌هایی مشخص است که مفهوم خاصی را اضافه می‌کند. تصویر ۲-۱۰ فرآیند محاسبه شباهت را نمایش می‌دهد [۱۳۶]. روش‌های پیشین با استفاده از داده‌های آموزش، پارامترهای بهینه روش را، محاسبه می‌کردند ولی در اینجا، محتوای اطلاعات جمله‌ها بدون الگوریتم یادگیرنده برای محاسبه شباهت استفاده شده است. محتوای اطلاعات جمله‌ها می‌توانند از پایگاه دانش واژگان و پیکره محاسبه شوند و در پی آن «اشتراک محتوای اطلاعات» بین دو جمله حساب شود. در انتها شباهت از نسبت بین اشتراک محتوای اطلاعات و اجتماع آن مشتق می‌شود. در زیر این فرآیند با جزئیات توضیح داده شده است.

در تصویر ۱-۱۲ نمونه‌ای از شبکه معنایی واژگان آورده شده است و در تصویر ۲-۱۱ روابط اجتماع و اشتراک بین مفاهیم مشخص می‌باشد. از دیدگاه نظریه اطلاعات، اطلاعات برای حذف عدم قطعیت مورد استفاده قرار می‌گیرد و هرچه مفهوم مورد بررسی در سطح بالاتری باشد حاوی اطلاعات کمتری است. برای مثال، مفهومی که اطلاعات «male» و «female» را ارائه می‌دهد قطعاً اطلاعات «person» را هم در خود دارد. به عبارتی دیگر از دید روابط اطلاعات معنایی، «person» در داخل هردو گنجانده شده است. بنابراین می‌توان اطلاعات ارتباط معنایی را به شکلی مشابه با تصویر ۲-۱۱ محاسبه کرد.

۱- اشتراک محتوای اطلاعات بین مفهوم‌ها



تصویر ۲-۱۰: فرآیند ارزیابی شباهت در روش CTS [۵۴]



تصویر ۲-۱۱: (سمت راست) احاطه‌ی محتوای اطلاعات و (سمت چپ) احاطه‌ی مفهوم‌های مرتبط [۵۴]

برای محاسبه « محتوای اطلاعات» از روش رِزنیک و طبق فرمول ۲-۲۳ الهام گرفته شده است.

$$IC(c) = -log(P(c)) \qquad (۲۳-۲)$$

که در آن $p(c)$ بسامد رخداد مفهوم c است. محتوای اطلاعات یک واژه از احتمال آن در یک پیکره زبانی مشتق شده است. محاسبه محتوای اطلاعات از این روش بسیار مناسب است زیرا که هرچه احتمال رخداد افزایش پیدا می‌کند، محتوای اطلاعات واژه کاهش پیدا می‌کند؛ بنابراین هرچه مفهوم انتزاعی‌تر باشد محتوای اطلاعات آن کمتر است. از این معیار برای محاسبه بار اطلاعاتی یک مفهوم استفاده

می‌شود. اشتراک چند مفهوم با فرمول ۲-۲۴ به دست می‌آید.

$$commonIC(c_1, c_2, ..., c_n) = IC(c_1 \cap ... \cap c_n) = max_{c \in subsum(c_1,...,c_n)}[-log(p(c))] \qquad (۲۴-۲)$$

که در آن $subsum(c_1, ..., c_n)$ به لیست زیرمجموعه‌های مشترک بین $(c_1, ..., c_n)$ اشاره دارد، زیرا که ممکن است چندین والد وجود داشته باشد. برای مثال بین مفهوم «مرد» و «خانم» زیرمجموعه‌ی مشترک «شخص» است. به صورت خاص اگر که فقط یک مفهوم وجود داشته باشد فرمول ۲-۲۴ تبدیل به ۲-۲۵ خواهد شد.

$$commonIC(c_1) = max_{c \in c_1)}[-log(p(c))] = IC(c_1) \qquad (۲۵-۲)$$

۲- محاسبه شباهت بین جمله‌ها

اگر دو جمله را مطابق فرمول ۲-۲۶ تعریف کنیم.

$$\begin{aligned} s_a &= \{c_i^a | i = 1, 2, ..., n_a; n_a = |s_a|\} \\ s_b &= \{c_i^b | i = 1, 2, ..., n_b; n_b = |s_b|\} \end{aligned} \qquad (۲۶-۲)$$

بنابراین محتوای اطلاعات کل جمله اول به صورت ۲-۲۷ محاسبه می‌شود.

$$IC(s_a) = IC[\cup_{i=1}^{n} c_i^a] = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < ... < i_k \le n} IC(c_{i1}^a \cap ... \cap c_{ik}^a) \qquad (۲۷-۲)$$

طبق این فرمول، برای مثال در جمله زیر:

I am a good and important person

بعد از حذف ایست‌واژه‌ها به ترتیب ابتدا اشتراک «good» و «important» پیدا می‌شود، سپس در فرمول با اشتراک «good»، «important» و «person» جمع خواهد شد. اجتماع محتوای اطلاعات دو جمله با فرمول ۲-۲۸ محاسبه می‌شود.

$$IC(s_a \cup s_b) = IC[\cup_{t=a,b}[\cup_{i=1}^{n_t} c_i^t]] = \sum_{k=1}^{n}(-1)^{k-1} \sum_{1\le i_1<\ldots<i_k\le n} IC(c_{i1}^a \cap \ldots \cap c_{ik}^b) \qquad (۲۸-۲)$$

و اشتراک بین دو جمله با استفاده از «اجتماع» در فرمول ۲-۲۹ محاسبه می‌شود. در روش نگاشت مفهوم‌ها، فرمول نهایی شباهت مطابق با ۲-۳۰ است.

$$IC(s_a \cap s_b) = IC(s_a) + IC(s_b) - IC(s_a \cup s_b) \qquad (۲۹-۲)$$

$$sim(s_a, s_b) = \frac{IC(s_a \cap s_b)}{IC(s_a \cup s_b)} \qquad (۳۰-۲)$$

اما در روش هوآنگ و همکاران، شباهت بین موجودیت‌ها باید طبق فرمول ۲-۳۱ محاسبه شود [۵۴].

$$oovIC(NEs) = \sum_{ne\in N} IC(ne) \qquad (۳۱-۲)$$

برای بهبود نتایج روش CTS، برای هریک از مفهوم‌های موجود در جمله، وابسته به نوع واژه متناظر آن (نام، عدد، نام موجودیت) وزن مشخصی در نظر گرفته می‌شود و فرمول محتوای اطلاعات وزنی جمله

بعد به فرمول ۲-۳۲ تغییر می‌کند.

$$wIC(S) = IC(S) + \sum_{i=1}^{n}[IC(c_i) \times (weight(w_i - 1)] \quad (۳۲-۲)$$

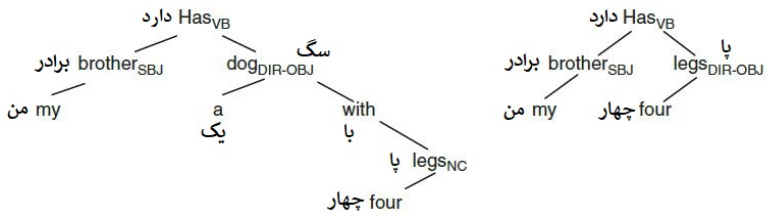
نام موجودیت‌ها نیز به دودسته «شناخته‌شده (cNE)» و «همگی (aNE)» تقسیم شده است و برای موجودیت‌های هر دسته وزن متفاوتی در نظر گرفته می‌شود تا فرمول کلی ۲-۳۱ به ۲-۳۳ تغییر کند و روش نهایی محاسبه شباهت به فرمول ۲-۳۴ تغییر می‌کند.

$$woovIC(NEs) = \sum_{ne \in NEs} [weight(ne) \times oovIC(ne)] \quad (۳۳-۲)$$

$$sim_{cts}(s_a, s_b) = \frac{wIC(s_a \cap s_b) + woovIC(cNEs)}{wIC(s_a \cup s_b) + woovIC(aNEs)} \quad (۳۴-۲)$$

SymSS ۷-۳-۲

روش SyMSS بر این مفهوم استوار است که معنی یک جمله را، نه‌تنها معانی واژگان مجزای آن تشکیل می‌دهد، بلکه به روش ساختاری که باهم ترکیب می‌شوند نیز وابسته است [۹۵]. بنابراین، SyMSS اطلاعات معنایی و ساختاری را برای محاسبه شباهت دو جمله ترکیب می‌کند. اطلاعات معنایی از «پایگاه داده لغوی شبکه واژگان» به دست می‌آید [۶۸]. اطلاعات ساختاری از طریق یک فرآیند تجزیه عمیق به دست می‌آید که عبارت‌ها (گروهی از واژگان که باهم یک معنی واحد می‌دهند) را در جمله پیدا می‌کند. با این اطلاعات، شباهت معنایی میان مفاهیمی که نقش نحوی مشابهی را ایفا می‌کنند، اندازه‌گیری شده است. با توجه به پژوهش‌های اخیر در ارتباط با نحوه اهمیت دادن انسان‌ها به نقش‌های



تصویر ۲-۱۲: درخت تجزیه جمله [۹۵]

متفاوت ساختاری، وزن‌های مختلفی تخصیص داده شده است. این پژوهش به نقش مهم اطلاعات ساختاری جمله در محاسبه شباهت اشاره دارد و بحث می‌کند که مشکل اصلی روش‌های پیشین و آنالیز معنای نهفته در استفاده نکردن از اطلاعات ساختاری جمله برای محاسبه شباهت است. برای مثال دو جمله زیر از دید آنالیز معنای نهفته یکسان هستند در حالی که این‌طور نیست و همچنین امکان نفی واژگان در نظر گرفته نمی‌شود (واژگانی مانند not و no).

- The dog chased the man
- The man chased the dog

دو جمله بالا، به ترتیب «سگ مرد را دنبال کرد» و «مرد سگ را دنبال کرد» ترجمه می‌شوند. همه‌ی مترادف‌های هر واژه برای محاسبه شباهت استفاده می‌شود. در صورتی که نسبت به جمله‌ی اول، نقش‌واژه متناظر در جمله‌ی دوم وجود نداشته باشد، مقدار ارزیابی شباهت، به مقدار ثابت و کوچکی جریمه می‌شود. در تصویر ۲-۱۲ برای دو جمله زیر درخت تجزیه جمله‌ها نمایش داده شده است.

- My brother has a dog with four legs
- My brother has four legs

فرض بر این‌که جمله s_1 از n عبارت تشکیل شده باشد و حاوی $h_1, ..., h_n$ جزء باشد، تابع محاسبه شباهت مطابق فرمول ۲-۳۵ است.

$$sim_{SymSS}(s_1, s_2) = \frac{1}{n}\sum_{i=1}^{n} sim_{path}(h_{1i}, h_{2i}) - c.PF \qquad (۳۵-۲)$$

$$sim_{path}(h_{1i}, h_{2i}) = \frac{1}{1 + length(h_{1i}, h_{2i})} \quad (۳۶-۲)$$

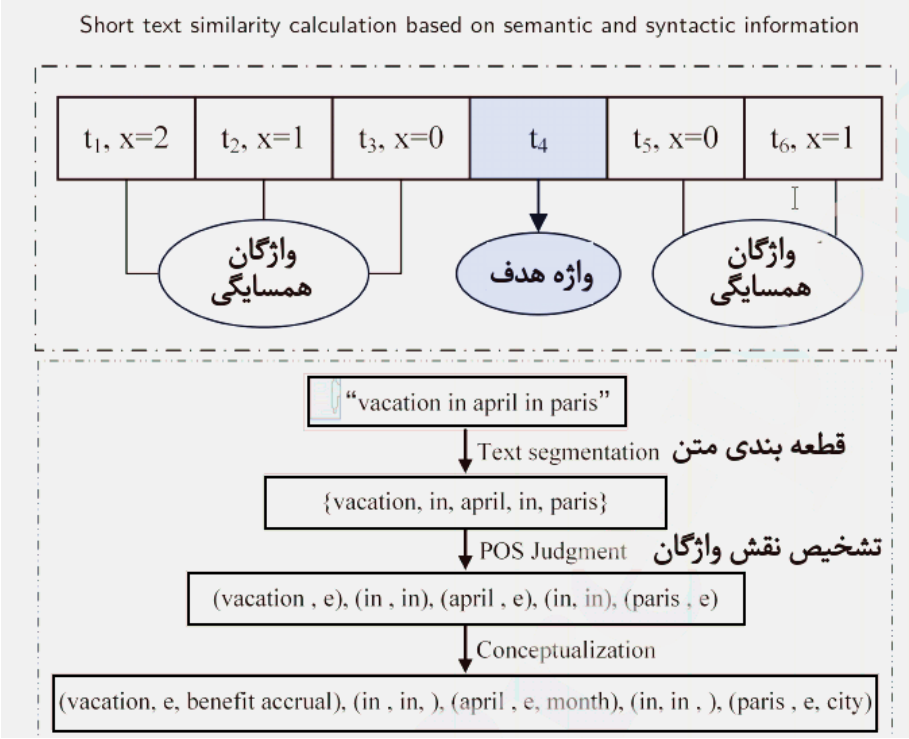
به این صورت اجزاء متناظر با یکدیگر مقایسه می‌شوند. c تعداد اجزاء متناظری است که یکسان نیستند و PF جریمه اآن‌ها است؛ به این معنی که اگر دو جزء متناظر یکسان نباشند، نمره ارزیابی شباهت کاهش می‌یابد. لازم به ذکر است که روش SymSS، از واژگانی که در شبکه‌ی واژگان نباشد چشم‌پوشی می‌کند.

از پژوهش‌های دیگر اخیر با رویکردی مشابه، می‌توان به روش یانگ و همکاران اشاره کرد [۱۴۰]. روش کلی این پژوهش در تصویر ۲-۱۳ نمایش داده شده است. ابتدا واژگان از جمله استخراج می‌شوند، سپس نقش‌واژه هرکدام تعیین می‌شود. در مرحله بعد باید مشخص شود که هر واژه بین تمامی معناهای ممکن به کدام‌یک مرتبط‌تر است (ابهام‌زدایی واژه). برای مثال، واژه «apple» ممکن است به معنای سیب و یا شرکت اپل باشد و باید بین این دو یکی انتخاب شوند. سپس با توجه به هر واژه در جمله اول و هر واژه در جمله دوم، بردار ویژگی برای دو جمله شکل می‌گیرد و فاصله کاسینوس به عنوان معیار ارزیابی فاصله بین دو جمله محاسبه می‌شود.

۲-۴ رویکردهای مبتنی بر شبکه‌های عصبی

۲-۴-۱ Word2Vec

مدل Word2vec تخمینی به‌منظور یادگیری بردارهای واژگان از متون است که از لحاظ پیچیدگی محاسباتی بسیار کارآمد می‌باشد [۸۶]. به بیان ساده‌تر در این مدل قرار است روابط بین واژگان را از موقعیت قرارگیری آن‌ها در متون استخراج کند. Word2vec به دو صورت کلی پیاده‌سازی می‌شود: مدل



تصویر ۲-۱۳: دیاگرام روش یانگ و همکاران [۱۴۰]

مجموعه واژگان پیوسته[14] و اسکیپ‌گرام[15].

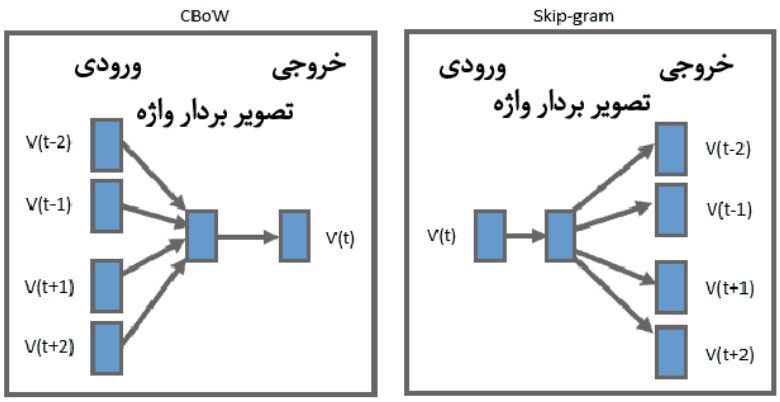
از لحاظ الگوریتمی این دو روش شبیه به هم هستند با این تفاوت که مدل مجموعه واژگان پیوسته واژه هدف را از روی واژگان همسایگی پیش‌بینی می‌کند، ولی اسکیپ‌گرام این فرآیند برعکس است.

برعکس کردن این چرخه دلخواه به نظر می‌رسد ولی از لحاظ آماری مدل پیوسته، تأثیر نرمی بر روی همه اطلاعات توزیعی دارد (با رفتاری شبیه به یک مشاهده بر روی کل متن) و درکل، این روش می‌تواند روشی مفید برای استفاده در مجموعه دادگان کوچک‌تر باشد. اما اسکیپ‌گرام با هر زوج «واژگان محتوا-واژگان هدف» به‌صورت یک مشاهده جدید رفتار می‌کند و در مجموعه دادگان بزرگ‌تر بهتر جواب می‌دهد.

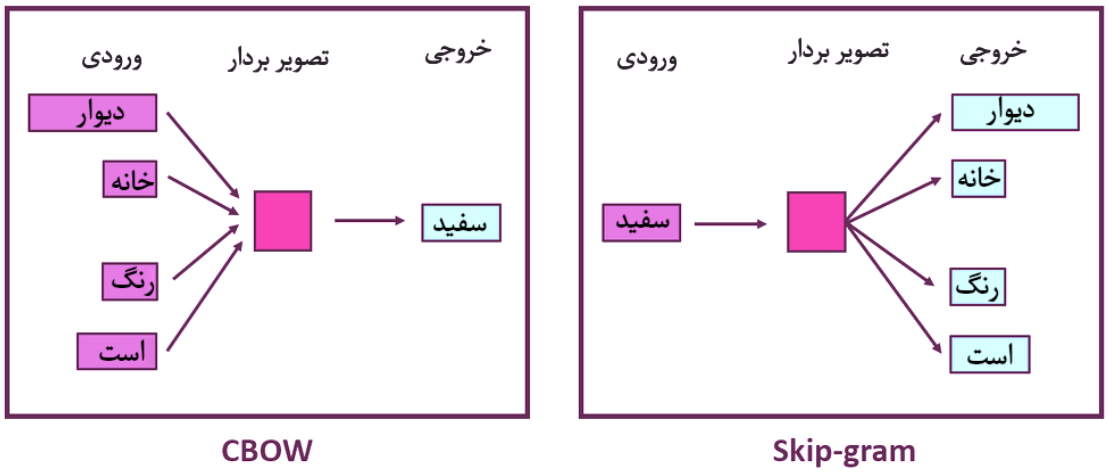
همان‌طور که در تصویر ۲-۱۴ مشخص است در Word2Vec سعی می‌شود تا بردار ویژگی واژگان طوری تولید شود که بتواند واژه هدف را پیش‌بینی کند. به عبارتی روی تمام پیکره متنی، الگوریتم اجرا می‌شود تا یک بردار عمومی برای هر واژه تولید شود. پس از آن، بردار واژه مستقل از جمله یا متنی که

[14] Continuous bag of words

[15] Skip-gram



تصویر ۲-۱۴: بردار ویژگی برای روش‌های Word2Vec [۱۴]

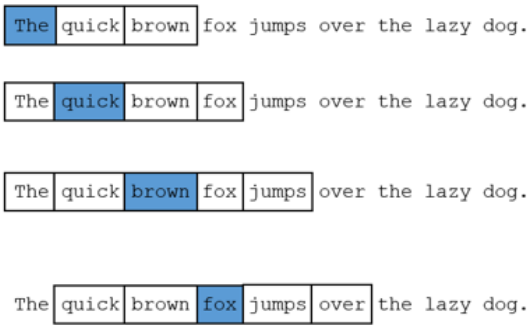


تصویر ۲-۱۵: مثال تولید بردار واژه با روش Word2Vec

در آن به کار می‌رود ثابت است. به این نوع روش‌های تولید بردار، روکیردهای مستقل از محتوا نیز گفته می‌شود. نمونه این فرآیند را می‌توان با مثال در تصویر ۲-۱۵ مشاهده کرد. یک پنجره شناور، روی متن حرکت می‌کند و واژه هدف توسط واژگان همسایه تخمین زده می‌شود، اندازه این پنجره می‌تواند متغیر باشد و وابسته به آن، بردار تولید شده، متفاوت خواهد بود.

Glove ۲-۴-۲

روش Glove یک الگوریتم یادگیری بدون ناظر و مستقل از محتوا، برای تولید نمایش‌های برداری واژگان است [۱۰۱]. آموزش روی ارقام هم‌رخدادی واژه به واژه که از روی یک پیکره بزرگ ضبط شده است، صورت می‌گیرد. توصیفگرهای نهایی زیرساختارهای خطی جالبی را در فضای بردار واژگان نمایش



تصویر ۲-۱۶: پنجره شناور برای تولید بردار واژه توسط روش Word2Vec [۷]

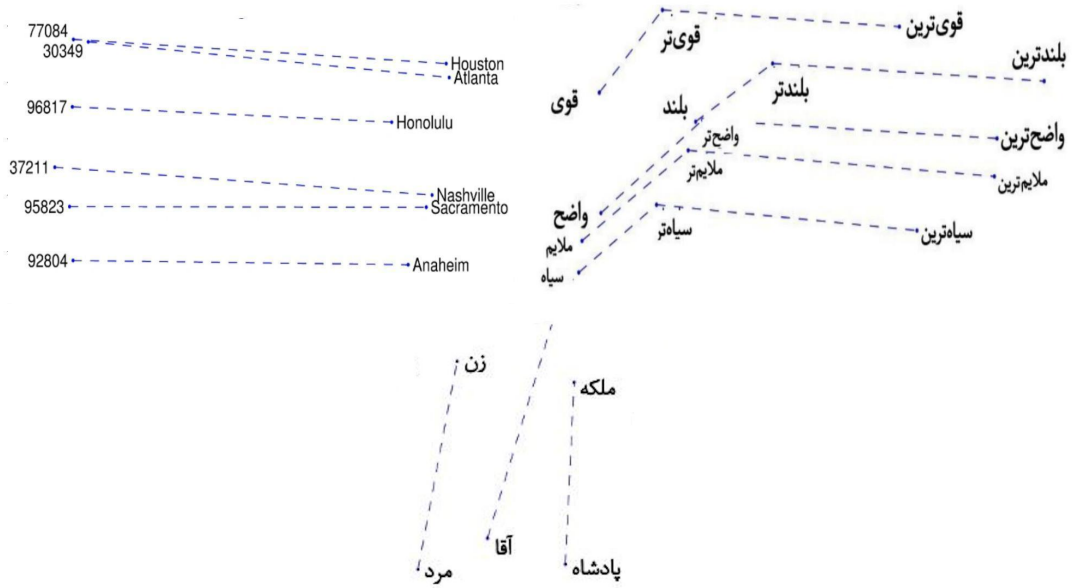
جدول ۲-۵: نمونه جدول اطلاعات محاسبه شده توسط Glove

k=fashion	k=water	k=gas	k=solid	احتمالات
$1/7 \times 10^{-5}$	3×10^{-3}	$6/6 \times 10_{-5}$	$1/9 \times 10^{-4}$	*p(k|solid)*
$1/8 \times 10^{-5}$	$2/2 \times 10^{-3}$	$7/8 \times 10^{-4}$	$2/2 \times 10^{-5}$	*p(k|gas)*
0/96	1/36	$8/5 \times 10^{-2}$	8/9	*p(k|solid) / p(k|gas)*

می‌دهند. این مدل، بر روی ماتریس غیر صفر هم‌رخدادی واژگان آموزش داده می‌شود که نیازمند یک‌بار جمع‌آوری آمار از روی پیکره متن است. برای یک پیکره متن بزرگ این مرحله می‌تواند بسیار طول بکشد و محاسبات سنگینی را نیازمند باشد، اما این مرحله فقط یک‌بار نیاز است انجام شود. سپس مراحل آموزش بسیار سریع‌تر هستند چراکه تعداد عناصر ماتریس غیر صفر به‌طور معمول بسیار کمتر از تعداد کل واژگان در پیکره است.

بینش اصلی در طراحی مدل این است که محاسبه احتمالات هم‌رخدادی واژگان می‌تواند یک نوع مدل معنایی کدگذاری کند. برای مثال احتمالات واژگان «یخ» و «بخار» را با واژگان دیگر متن در نظر بگیرید. در اینجا مقادیر احتمالات واقعی از یک پیکره متن حاوی شش میلیارد واژه آورده شده است.

هدف آموزشی Glove به این صورت است که بردار واژگان را به نحوی یاد بگیرد تا حاصل‌ضرب نظیر به نظیر واژگان برابر با لگاریتم احتمال هم‌رخدادی آن‌ها شود. همان‌طور که در جدول ۲-۵ دیده می‌شود، «یخ (ice)» بیشتر همراه با «جامد (solid)» رخ می‌دهد تا «گاز (gas)»؛ در حالی که «بخار» بیشتر با «گاز» رخ می‌دهد تا «جامد». هردو واژه به مقدار مشابهی با «آب» هم‌رخداد هستند در حالی که رابطه خاصی با «مُد(fashion)» ندارند. به این خاطر که ارقام نهایی دارای معنا می‌باشند، اطلاعات تولید شده می‌توانند توسط عملیات برداری هم مورد استفاده قرار بگیرند. به عنوان مثال تصویر ۲-۱۷



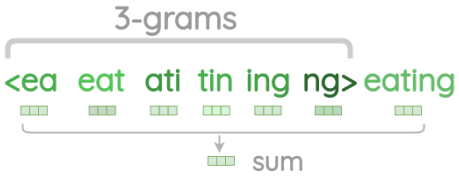
تصویر ۲-۱۷: روابط برداری بین واژگان توسط Glove [۱۰۱]

را مشاهده کنید.

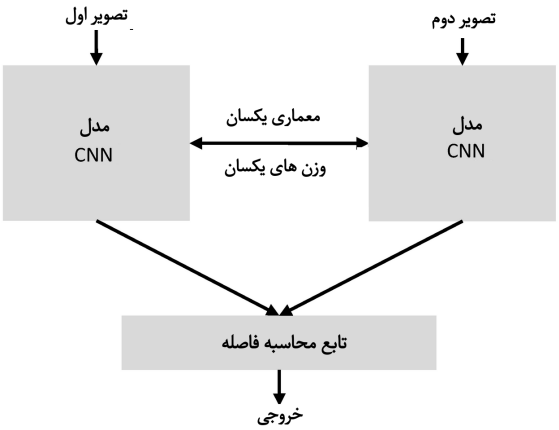
در تصویر سمت راست، می‌شود روابط بین صفات مقایسه‌ای و عالی را، با صفت ریشه مشاهده کرد. در واقع، فاصله بین «قوی» و «قوی‌تر» مانند فاصله بین «سیاه» و «سیاه‌تر» است. همچنین واژگان مشابه به هم در فضای برداری نزدیک‌تر هستند. برای مثال «واضح» و «بلند» به هم نزدیک‌تر هستند، تا فاصله بین «قوی» و «سیاه» که می‌توان گفت نامرتبط هستند. این روابط خطی در تصویر سمت چپ هم دیده می‌شود، چنانکه شماره کد پستی هر شهر، به شهر متناظر نگاشت شده است. این روابط خطی در تصویر پایینی هم دیده می‌شود. به عبارتی فاصله «مرد» با «زن» مساوی فاصله «پادشاه» با «ملکه» است. به عبارتی می‌توان عنوان متناظر برای «پادشاه» را با عملیات برداری خطی و با دانستن فاصله «مرد» یا «زن» به دست آورد.

FastText ۳-۴-۲

با توجه حجم عظیم داده‌های تولید شده توسط کاربران فیس‌بوک، این شرکت چالش سختی را برای مدیریت بهتر کاربرانش داشت و روش‌های یاد شده مشکل فیس‌بوک را حل نمی‌کرد. FastText یک کتابخانه تولید شده به‌منظور توصیف واژگان و کلاسه‌بندی آن‌ها است. رویکرد FastText (که افزونه‌ای



تصویر ۲-۱۸: نحوه تشکیل بردار توصیف واژه توسط FastText [۱۶]



تصویر ۲-۱۹: مدل شبکه‌های موازی برای تشخیص یکسان بودن تصاویر [۱۶]

از مدل Word2Vec است) ، هر واژه را، به عنوان ترکیبی از n-gramهای کاراکترهای تشکیل‌دهنده آن در نظر می‌گیرد. بنابراین، بردار یک واژه، از مجموع حالت‌های مختلف n-gram آن ساخته شده است [۱۶]. برای مثال واژه «apple» توسط جمع مجموعه نویسه‌های زیر توصیف می‌شود. از مزایای این روش، می‌توان به توصیف واژگانی که در پیکره آموزش وجود ندارند، اشاره کرد.

”<ap”, ”app”, ”appl”, ”apple”, ”apple>”, ”ppl”, ”pple”, ”pple>”, ”ple”, ”ple>”, ”le>”

همان‌طور که در تصویر ۲-۱۸ آمده است یک واژه به مجموعه‌های سه‌نویسه‌ای تبدیل شده، که مجموع بردارهای آن‌ها، توصیف واژه را می‌سازند.

۲-۴-۴ ROT

در یک روش ارزیابی شباهت جمله با عنوان «آگاه از ساختار»، وانگ و همکاران از الگوریتم حمل‌ونقل بهینه بازگشتی[۱۶] برای محاسبه شباهت جمله استفاده کردند [۱۳۰]. این الگوریتم، هر جمله را به یک ساختار درختی تجزیه می‌کند و واژگان متناظر (هم‌تراز) بین دو جمله را، به هم تطبیق می‌دهد.

[16]Recursive optimal transport algorithm

هر زیرشاخه در این درخت، توسط جمع وزنی بردار واژگان آن توصیف می‌شود. بنابراین هر جمله توسط درخت آن و مجموع وزنی بردارهای زیردرخت‌ها توصیف می‌شود. برای بردارهای اولیه واژگان، توصیفگرهایی مانند Word2Vec مورد استفاده قرار می‌گیرند؛ نمره نهایی شباهت نیز مبتنی بر متضاد فاصله کاسینوسی بین دو بردار است.

۲-۵ تعمیم بردارهای واژگان به بردار واحد جمله

بدین منظور می‌توان دو رویکرد را در نظر گرفت. در روش باناظر، با استفاده از پایگاه داده‌های آموزشی، طوری بردار ویژگی جمله آموزش داده می‌شود که جمله‌های شبیه‌تر از نظر برداری به هم نزدیک‌تر شوند. در روش بدون ناظر، بردار ویژگی برای جمله صرف‌نظر از وجود داده‌های آموزشی تشکیل می‌شود. هر دو حوزه از حوزه‌های بسیار مهم پژوهشی هستند و نمونه‌هایی از آن‌ها را در ادامه بررسی می‌کنیم.

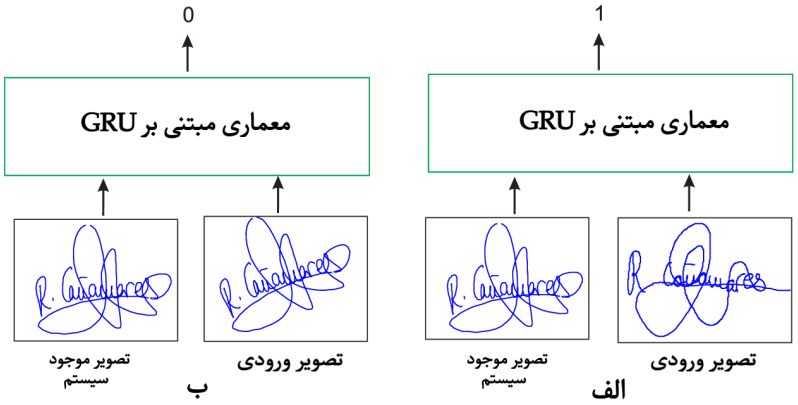
۲-۵-۱ Siamese

شبکه‌های سیامز[17]، ابتدا برای ارزیابی شباهت زوج تصاویر طراحی شدند. این شبکه‌ها دو جفت CNN، در طرف «هر تصویر مورد مقایسه» دارند (شکل ۲-۱۹). داده‌های ورودی شامل دو تصویر هستند و خروجی، تشخیص یکسان بودن یا نبودن اشیاء موجود در تصویر است. البته می‌توان برای اضافه کردن حافظه به شبکه، حافظه‌های همگشتی را هم به معادله اضافه کرد [۱۴۱].

نمونه‌ای از مثال کاربردی این شبکه در تصویر ۲-۲۰ شناسایی می‌کند که دو امضا مربوط به یک شخص هستند یا خیر. بنابراین کارکرد آن می‌توان تشخیص جعل امضا و یا تشخیص هویت اعضاء باشد [۱۲۳]. همان‌طور که در خروجی شبکه دیده می‌شود سیستم با نمایش یک یا صفر، اصل بودن امضا را مشخص می‌کند ضمن اینکه در این سیستم از معماری شبکه‌های همگشتی استفاده شده است.

هرچند که برای پردازش متن، امکان پیاده‌سازی همین معماری وجود دارد اما بهترین کارایی توسط سیستم‌های همگشتی به‌دست‌آمده است. نیاز به حافظه و یادسپاری الگوهای مهم در طراحی شبکه‌های

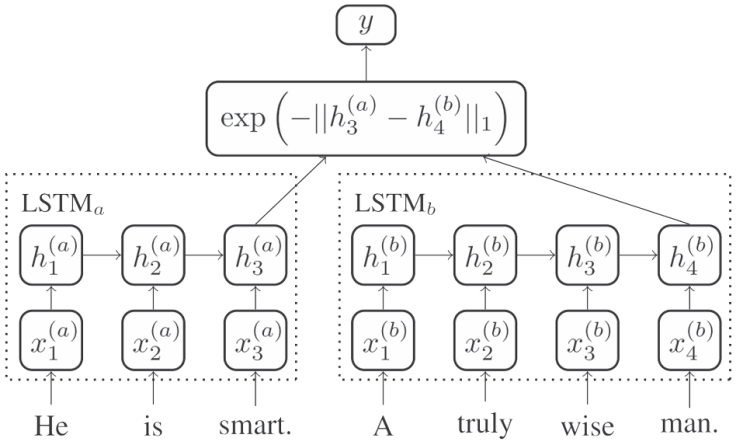
[17] Siamese



تصویر ۲-۲۰: تشخیص جعل امضا با استفاده از شبکه‌های همگشتی (الف) امضای جعلی و (ب) امضای اصل [۱۲۳]

Siamese برای متن یک باید است [۹۰]. شبکه Siamese استاندارد با حافظه LSTM مطابق تصویر ۲-۲۱ پیاده‌سازی می‌شود. هدف از طراحی این سیستم، تولید خروجی مقدار پیوسته‌ای است که توسط داوران انسانی به عنوان مقدار شباهت به جفت جملات تخصیص داده شده است. بنابراین، یک شبکه رگرسیون طراحی می‌شود، در انتهای آن فقط یک نورون سیگموید وجود دارد و مقداری در بازه صفر تا یک تولید می‌کند. مقدار یک به «مساوی بودن معنای دو جمله» و عدد صفر به «کاملاً متفاوت بودن دو جمله» اشاره دارد. در یکی از پژوهش‌های اخیر نیز، ابتدا درخت نقش‌واژگان جمله تشکیل شده است، سپس بردارهای هر واژه با توجه به «محل قرارگیری آن‌ها در درخت تجزیه» به یک شبکه عصبی عمیق ارسال می‌شوند [۱۱۲].

در بعضی پایگاه‌های داده ممکن است برای هر زوج‌جمله، فقط یکی از دو خروجی مدنظر باشد. به عبارتی فقط اینکه آیا دو جمله مساوی هستند یا خیر. در بحث ارزیابی میزان ربط جمله، بررسی می‌شود که آیا دو جمله صرف‌نظر از شباهت، به هم مرتبط هستند یا خیر. دو جمله ممکن است متضاد یا مساوی باشند ولی مقدار ربط مشابهی داشته باشند. البته در بحث استلزام متن، ارتباط جهت‌دار بین تکه‌های متن وجود دارد. به عبارتی یک متن، به عنوان فرضیه وجود دارد و در مقابل آن، متنی وجود دارد که باید مشخص شود متن فرضیه را تأیید یا رد می‌کند. به عبارتی سیستم باید توانایی درک متن داشته باشد.



تصویر ۲-۲۱: شبکه سیامز برای ارزیابی شباهت زوج‌جملات [۹۰]

۲-۵-۲ مبتنی بر توجه به نویسه‌های واژگان

در مدل توجه به نویسه‌ها[18]، پژوهشگران فهرستی از تمامی حالت نویسه‌های ممکن برای هر جمله را می‌سازند [۷۹]. سپس ماتریس سنجش[19]، با توجه به‌تمامی زوج-n-نویسه‌های ممکن بین هردو جمله، ساخته می‌شود. این باعث شده است تا نتایج نسبت به حالت بدون تکنیک توجه روی داده‌های SICK بهبود یابد.

۲-۵-۳ توجه به واژگان همسایگی

در روش توجه به واژگان همسایگی[20]، چندین پنجره به مرکزیت یک واژه در جمله تشکیل می‌شوند [۵۵]. به این ترتیب، میزان اهمیت هر واژه نسبت به پنجره‌های مختلف سنجیده می‌شود. دلیل این‌کار امکان شناسایی و جمع نمودن چند واژه است برای زمانی که یک عبارت را تشکیل می‌دهند و ممکن است به تنهایی معنای مفیدی نداشته باشند. برای سازو کار «تکنیک توجه»، تمامی توصیف واژگان به هم متصل می‌شوند و سپس یک کانولوشن یک‌بُعدی روی پنجره‌های مختلف اعمال می‌شود، سپس در عملیات بیشینه تجمعی[21]، از بردارهای مختلف ویژگی‌های مهم استخراج و به یک LSTM درختی ارسال

[18] n-gram attention

[19] Attention

[20] Window-based attention

[21] Max-Pooling

می‌شوند. در یک معماری مبتنی بر Tree-LSTM یا LSTM درختی، شبکه، وظیفه تولید بردار توصیفگر جمله را دارد [۱۱۹]. تفاوت آن با معماری LSTM معمولی در آن است که حالا حافظه LSTM در زمان جاری به مقادیر خود در هر گام قبلی دسترسی دارد و بنابراین می‌تواند از توصیف‌های فراهم‌شده در گام‌های قبلی استفاده کند.

۲-۵-۴ Disan

در روش Disan از سنجش چندبُعدی استفاده می‌شود [۱۱۳]. در حالت معمولی، سنجش همسایگی به ازای هر توکن (واژه یا عبارت) محاسبه می‌شود اما در این روش، برای هر بُعد توصیفگر واژه‌ها، تکنیک توجه به صورت مستقل محاسبه می‌شود؛ به صورتی که برای هر ویژگی بردار، نمره جدایی محاسبه شده است. در این پژوهش ادعا شده است که این امر به تشخیص معنای صحیح واژگان چندمعنایی با توجه به متنی که در آن به کار رفته‌اند کمک می‌کند.

بیشتر روش‌های اخیر، با اصلاح و اضافه کردن بخش کوچکی به رویکردهای پیشنهادی قبل‌تر، بهبودهای جزئی کسب کردند، به عنوان مثال بیشتر روش‌ها از شبکه‌هایی CNN یا LSTM استفاده می‌کنند. بیشتر پژوهش‌ها از نوعی تکنیک توجه نیز برای تشخیص واژگان کلیدی جمله استفاده می‌کنند. در این پژوهش ما روشی را پیشنهاد می‌دهیم که از توصیفگرهای متفاوت برای واژگان استفاده می‌کند. به عبارتی هر توصیفگر، دید متفاوتی به واژه را برای شبکه عصبی فراهم می‌کند.

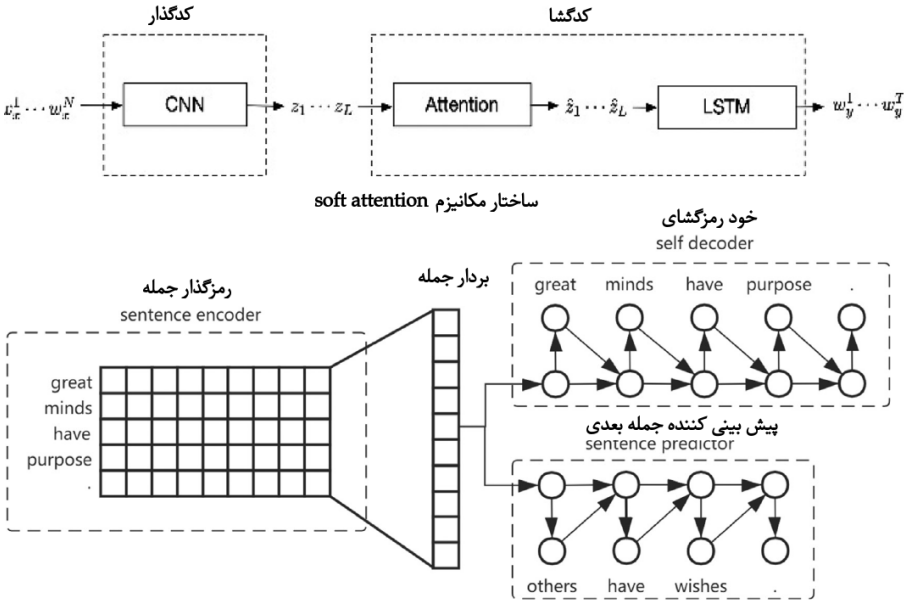
۲-۵-۵ روش کدگذار-کدگشای CNN-LSTM

در اینجا تعدادی از روش‌های توصیف جملات[۲۲] را بررسی می‌کنیم. در پژوهشی توسط فو وهمکاران، از تلفیق دو نورون CNN و LSTM به‌منظور ارزیابی شباهت استفاده شده است (شکل ۲-۲۲). ابتدا توصیفگرهای واژگان[۲۳] در ماتریسی قرار می‌گیرند به صورتی که در هر ردیف بردار توصیف یک واژه باشد. در لایه دوم چندین پنجره کانولوشن بر روی ماتریس توصیفگر اعمال می‌شود تا ویژگی‌های مؤثر

[22] Sentence representation

[23] Word embeddings

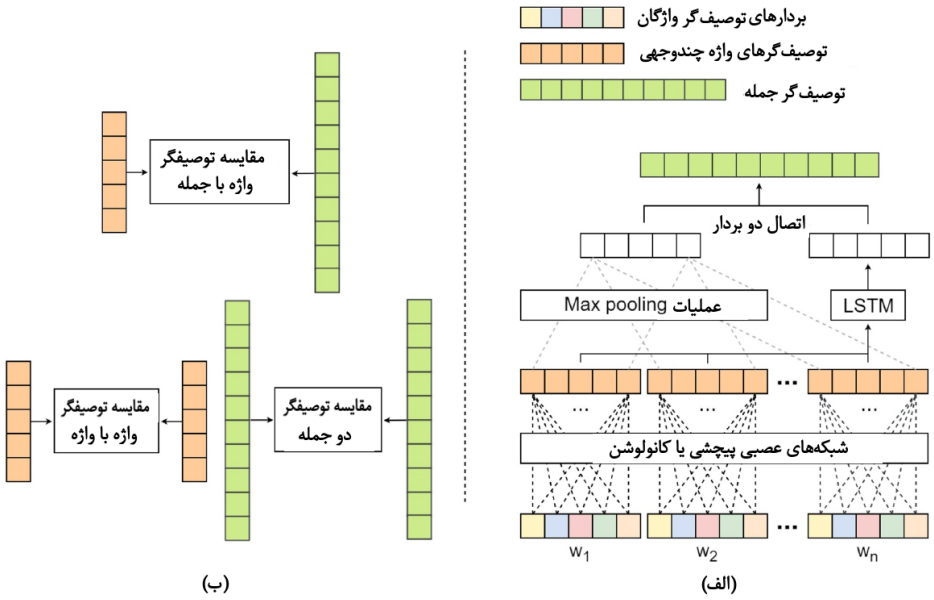
از بردارها استخراج شوند و واژگان کلیدی تشخیص داده شوند. این مرحله را می‌توان گام استخراج ویژگی یا کدگذاری نامید. در مرحله بعد دو نوع ساختار شبکه داریم؛ یک نوع آن اطلاعات دریافت شده توسط CNN را با استفاده از حافظه‌های LSTM دوباره بازسازی می‌کند. ساختار دیگر شبکه با همان معماری، اما واژگان جمله دوم را پیش‌بینی می‌کند. بعد از مرحله آموزش، برای برداری که بعد از لایه CNN تولید شده است به عنوان بردار ویژگی جمله استفاده می‌شود. متضاد فاصله بین دو بردار جمله را می‌توان به عنوان معیار ارزیابی شباهت استفاده کرد. حال این معیار ممکن است فاصله اقلیدسی، کاسینوسی و یا معیارهای دیگر باشد [۳۸].



تصویر ۲-۲۲: معماری شبکه کدگذار-کدگشای CNN-LSTM به‌منظور تولید بردار جمله [۳۸]

در پژوهشی مبتنی بر «وزن‌دهی بردار توصیف واژگان» [۸]، با تکیه بر اینکه توصیفگرها در فضای برداری حاوی روابط خطی هستند، از بسامدهای رخداد واژگان، به عنوان ضرایب جمع خطی توصیفگر آن‌ها استفاده شده است. این جمع خطی در نتیجه، بردار توصیفگر جمله را تولید می‌کند. همچنین در یک پژوهش اخیر، از توصیفگرهای نویسه و واژگان برای آموزش حافظه‌های LSTM، به‌منظور تولید بردار جمله نیز استفاده شده است [۱۴۴].

در پژوهشی دیگر، ابتدا پنجره‌های کانولوشن یک‌بعدی به هرکدام از توصیفگر واژگان اعمال می‌شود (شکل ۲-۲۳ (الف))؛ در مرحله بعد دو عملیات «انتقال به حافظه LSTM» و «انتخاب مقدار بیشینه از



تصویر ۲-۲۳: تشکیل بردار ویژگی جمله و محاسبه شباهت در روش MultiEmbedding [۱۲۱]

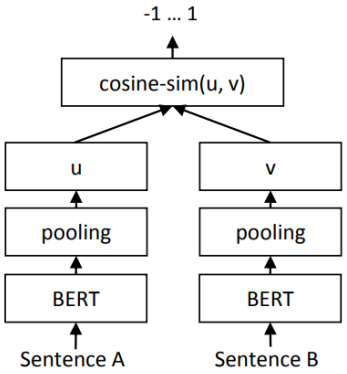
پنجره مقادیر جدید توصیفگر»، مستقل از هم انجام خواهد شد. در ادامه بردار تولید شده توسط هرکدام به همدیگر متصل می‌شود و توصیف جمله را می‌سازد. به‌منظور محاسبه شباهت بین دو جمله، سه رویکرد مقایسه‌ای، مطابق تصویر ۲-۲۳ (ب) انجام می‌شود [۱۲۱].

معیارهای فاصله شامل معیارهای کاسینوسی و میانگی تفریق مطلق هستند. در مقایسه بین دو توصیفگر چندوجهی، یک پرسپترون ساده استفاده می‌شود که مقدار شباهت نهایی را با توجه به ماتریس شباهت توصیفگرهای تمامی زوج‌واژگان تولید می‌کند. مقدار شباهت هر سه روش اشاره شده، بااتصال به همدیگر و تشکیل یک بردار، به شبکه پرسپترون دیگری ارسال می‌شوند تا مقدار نهایی شباهت را تولید نماید.

۲-۵-۶ توصیف‌گرهای مبتنی بر محتوا

در توصیف‌گر وابسته به محتوا، بردار توصیف‌گر واژه، وابسته به متنی که در آن به کار رفته است تغییر می‌کند. این روش یک اصلاحیه نسبت به شبکه BERT[۲۴] از پیش آموزش دیده است. این رویکرد، با استفاده از معماری Siamese (آنچه در شکل ۲-۲۴ نمایش داده شده است) یک توصیف برای جمله

[24]Bidirectional Encoder Representations from Transformers



تصویر ۲-۲۴: تشکیل بردار ویژگی جمله و محاسبه شباهت در روش Sentence-BERT[۱۰۹]

می‌سازد که در نتیجه آن، می‌توان با معیار کاسینوسی شباهت زوج‌جمله‌ها را محاسبه کرد [۱۰۹]. سیستم BERT نتیجه یک پژوهش اخیر توسط تیم گوگل می‌باشد و توانسته در بسیاری از کاربردهای پردازش متن، نتایج بهتری را نسبت به رقبا کسب کند [۳۱]. این نوع شبکه، توصیف واژگان را با توجه به (۱) محتوای دربرگیرنده آن در دو طرف متن بعلاوه (۲) تمام لایه‌های شبکه، می‌سازد. در آموزش، از تکنیک ماسک‌گذاری روی واژگان جمله استفاده شده است. برای مثال، ۱۵ درصد از آن‌ها حذف می‌شوند و شبکه باید بتواند با توجه به سایر واژگان در بافت متن، توصیف جمله را تخمین بزند. در بسیاری از این نوع سیستم‌ها، یک زیرسیستم نیز وجود دارد که در آن جمله بعدی یا مرتبط بودن یک متن دیگر به جمله ورودی، پیش‌بینی می‌شود. فرضیه آن است که افزودن این زیرسیستم، باعث بهبود نتایج و تولید توصیف‌گرهای بهتر برای جمله می‌شود اما پژوهش [۱۰۹] به این نتیجه رسیده است که حداقل روی سیستم BERT استفاده از این فن باعث کاهش کارایی سیستم در حل مسائل خاص می‌شود.

در پژوهشی، لیو و همکاران بحث می‌کنند که مدل اولیه BERT به اندازه کافی آموزش ندیده است و با آموزش بیشتر، می‌توان سیستم بهتری داشت. بنابراین، این بار آن‌ها هر جمله را ۱۰ بار به ۱۰ روش مختلف ماسک‌گذاری کردند و در نتیجه شبکه ۱۰ نوع مختلف از هر جمله را آموزش می‌بینند. آن‌ها این سیستم را RoBERTa[۲۵] نام‌گذاری کردند.

مدل XLM بر پایه BERT پیشنهاد شده است که محاسبه شباهت را برای چندین زبان توسعه می‌دهد [۷۱]. در آموزش این نوع سیستم، هر نمونه آموزش به دو زبان فراهم شده است و در نتیجه،

[25] A Robustly Optimized BERT Pretraining Approach

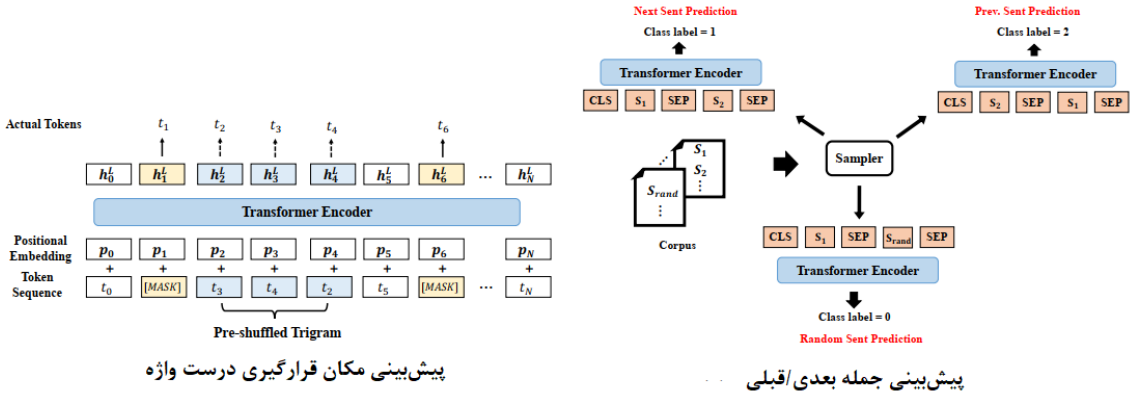
مدل می‌تواند در پیش‌بینی واژه ماسک شده از اطلاعات زبان دیگر استفاده کند. همچنین اطلاعات مکان قرارگیری واژگان نیز کدگذاری می‌شود و در نتیجه، مدل، روابط متناظر اجزاء دو زبان را نیز یاد می‌گیرد.

به دلیل (۱) منابع محدود آموزشی برای بهینه کردن شبکه برای یک مسئله خاص، (۲) پیچیدگی بالای مدل‌های پیش‌آموزش دیده و (۳) بهینه‌سازی بی‌اندازه این مدل‌ها، روی داده‌های آموزشی برای حل مسئله‌های خاص بیش‌برازش روی می‌دهد. بنابراین مدل روی داده‌های آزمون دیده نشده به خوبی تعمیم نمی‌یابد. مسائل خاص شامل موارد (اما نه محدود به) ارزیابی شباهت، تشخیص سرقت ادبی، پرسش و پاسخ خودکار می‌باشد. مدل SmartRoberta به اختصار Smart برای حل این مشکل آموزش خصمانه[۲۶] انجام می‌دهد. در این رویکرد، روی بخشی از بردارهای توصیف‌گر واژگان، مقادیر به صورت تصادفی تغییر می‌کند، اما در خروجی شبکه سعی می‌شود که خروجی با توجه به نوع مسئله، درست محاسبه شود. علاوه بر این، با روشی به نام بهینه سازی نقطه پروگزیمال Bregman [۲۷] یک تابع کمینه‌سازی وجود دارد که اجازه تغییرات ناگهانی در اندازه وزن‌های شبکه نمی‌دهد [۶۵]؛ به این صورت که تغییرات بزرگ در پارامترهای شبکه شامل جریمه می‌شوند.

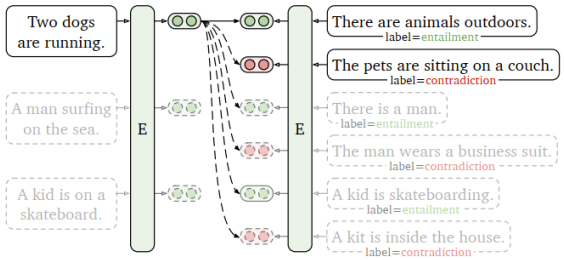
در روش StructBERT، اطلاعات ساختار قرارگیری واژگان جمله نیز در آموزش بردارهای توصیف‌گر در نظر گرفته می‌شود. این مدل بحث می‌کند که BERT معمولی نمی‌تواند اطلاعات ترتیب واژگان و وابستگی آن‌ها را به خوبی مدل کند. این رویکرد از معماری شبکه‌های مبدل دوطرفه چندلایه استفاده می‌کند. وقتی یک متن یا جفت‌جمله به شبکه داده می‌شود؛ تمامی دنباله واژگان در یک توکن فشرده می‌شود. بدین صورت هر توکن ورودی بر اساس واژه، مکان قرارگیری آن در جمله و متنی که در آن قرار گرفته است، توصیف می‌شود. این سیستم خود از دو زیر سیستم تشکیل می‌شود، بخش اول آن همانطور که در تصویر ۲-۲۵ آمده است، تلاش می‌کند تا با تغییر مکان قرارگیری واژگان جمله، مکان درست آن‌ها را در خروجی شبکه درست تخمین بزند. بخش دوم سیستم یک جفت جمله ورودی (s_1, s_2) به سیستم می‌دهد و مدل باید تخمین بزند که s_2 کدام‌یک از جمله قبلی، بعدی و یا یک جمله تصادفی از متن است. [۱۲۹].

[26] Adverserial training

[27] Bregman Proximal Point Optimization



تصویر ۲-۲۵: تصویر دو زیربخش شبکه StructBERT [۱۲۹]



تصویر ۲-۲۶: تصویر سیستم باناظر SimCSE [۳۹]

در شبکه نوع MNet-Sim، ابتدا از توصیف‌گرهای مبتنی بر BERT برای توصیف جمله استفاده شده است. در ادامه، چندین نوع شبکه تعریف شده که هرکدام با یک معیار فاصله مجزا، از جمله: اقلیدسی، کاسینوسی و معیار جاکارد شباهت را محاسبه می‌کنند. این شبکه‌ها به عنوان لایه‌های تشکیل دهنده شبکه هستند و در لایه نهایی مقدار خروجی شباهت زوج‌جمله محاسبه می‌شود [۶۴].

در روش SimCSE، برای توصیف‌گر جمله از روش BERT یا RoBERTa استفاده می‌شود [۷۸]. مسئله استنتاج جمله در این روش به جای پیش‌فرض‌های (استلزام/تناقض/خنثی) به دو حالت مثبت (مرتبط به هم) و منفی (نامرتبط یا متناقض) مدل می‌شود. نمایی از این مسئله در تصویر ۲-۲۶ مشخص شده است. بدین شکل، بردار توصیف جمله به‌روزرسانی می‌شود تا بتواند برچسب بین زوج‌جمله را درست تخمین بزند [۳۹].

۲-۵-۷ پیشنهادات پژوهشی

پایگاه دانش Yago، شبکه‌ی واژگانی می‌باشد که اخیراً معرفی شده[۱۴۶] است و حاوی اطلاعات شبکه‌ی واژگان بعلاوه ویکی‌پدیا و GeoNames [۲۸] است. با استفاده از پایگاه GeoNames، می‌توان نام موجودیت‌های بسیار بیشتری نسبت به آنچه ذر شبکه واژگان WordNet هست را مورد مقایسه قرار داد. در این پژوهش، با استفاده از محاسبه‌های شبکه واژگان Yago با معرفی یک معیار ترکیبی از وابستگی چند معیار فاصله، حالت دقیق‌تری برای محاسبه معیار شباهت تعریف می‌شود؛ هرچند که در پژوهشی توسط هو و همکاران نیز از اطلاعات ویکی‌پدیا برای محاسبات شباهت بین اسناد استفاده شده است [۵۲، ۱۰۴]. بنابراین ما در محاسبه «فاصله کوتاه‌ترین مسیر وزن‌دهی شده» (صفحه ۳۷) بین دو واژه از پایگاه دانش Yago استفاده می‌کنیم.

پایگاه داده Google ngram حاوی اطلاعات بسامد واژگان، پردازش شده از میلیون‌ها کتاب است [۱۹]. با استفاده از آن می‌توان به اطلاعاتی ازجمله بسامد تکرار دو واژه کنار هم یا تعداد کتاب‌هایی که واژه‌ای خاص در آن تکرار شده است، دسترسی داشت. از این پایگاه نیز برای محاسبات آماری شباهت بین دو واژه استفاده شده است. می‌توان واژگان اطراف دو واژه مورد مقایسه را لیست کرد و بر اساس بسامد تکرار واژگان همسایه، شباهت را محاسبه کرد.

پایگاه داده‌ی واژگان جایگزین (PPDB) نیز در این پژوهش استفاده شده است [۹۹]. این پایگاه داده حاوی میلیون‌ها زوج‌واژه مترادف از ۱۶ زبان مختلف است. استفاده از این پایگاه داده در تلفیق با روش‌های مبتنی بر شبکه‌ی واژگان، نتایج را بهبود داده است.

محاسبه شباهت زوج‌جمله‌ها از طریق مدل‌های عصبی پیش‌آموزش دیده هم مورد بررسی قرار گرفته است.

[28] Geonames (2021). Geographical Names [online]. webiste: https://www.geonames.org/ [accessed 2021 July 10]

نتیجه‌گیری روش موجک

در فصل بعدی، مدلی بر مبنای تبدیل موجک روی توصیف‌گرهای جمله پیشنهاد می‌شود که نتایج مدل پایه را بهبود می‌دهد و در فصل ۴ مورد آزمون قرار گرفته است.

۶-۲ جمع‌بندی

تعدادی از روش‌های پیشین، تأکید بالایی روی استفاده از معیارهای مبتنی بر شبکه واژگان برای محاسبه شباهت داشته‌اند. ویژگی مشترک بیشتر این روش‌ها در استفاده از یک نوع ماتریس شباهت واژه به واژه، برای محاسبه شباهت نهایی بین دو جمله است. روش‌های STASIS و LSS از این روش‌ها هستند که با استفاده از اطلاعات موجود در دو جمله و معیارهای شبکه واژگان، برای هر جمله یک بردار ویژگی می‌سازند و شباهت را با معیار کاسینوس بین دو جمله محاسبه می‌کنند. با وجود موفقیت بالای این روش‌ها، هیچ‌کدام غیر از روش SymSS احتمال منفی بودن یک جمله را بررسی نمی‌کند (وجود واژگانی که معنی جمله را نفی می‌کنند). از بین روش‌های بررسی شده، تنها روش STS است که از معیارهای کاملاً آماری استفاده می‌کند. روش‌های مبتنی بر شبکه عصبی، از ساختاری پیچیده‌تر برای توصیف هر جمله به صورت یک بردار، استفاده می‌کنند. این نوع، بهترین نتایج را در زمینه ارزیابی شباهت زوج‌جمله‌ها کسب کرده‌اند و افزودن «فن توجه» به معماری آن‌ها توانسته است نتایج را بهبود بخشد. در ادامه روش‌های پیشنهادی برای حل مسئله معرفی می‌شوند.

فصل ۳: روش‌های پیشنهادی

در این بخش، ابتدا روش‌های پیشنهادی برای حل مسئله‌ی شباهت متون کوتاه معرفی می‌شوند و در ادامه راهکارهای تلفیق این روش‌ها برای به دست آوردن بیشترین وابستگی با نمره‌های داوران انسانی تعریف می‌شوند. برای پیشنهاد راه‌حل از راهکاری مختلفی استفاده شده است. این راه‌حل‌ها مبتنی بر پایگاه دانش، اطلاعات آماری و یا عملیات تطبیق ساده رشته‌ها هستند. همچنین سه روش توصیف‌گر واژگان در قالب معماری Siamese برای ارزیابی شباهت زوج‌جمله‌ها مورد بررسی قرار گرفتند. این سه روش هرکدام از دید متفاوتی به روابط بین واژگان و بافت متن می‌پردازند. قبل از محاسبات ابتدا ایست‌واژه‌ها حذف می‌شوند و واژگان توسط ابزار استنفورد ریشه‌یابی می‌شوند [۸۱]. نوآوری‌های این رساله را می‌توان به صورت زیر تعریف کرد:

- از داده‌های Google ngram در محاسبات آماری استفاده شده است که اطلاعات بسامد دقیق‌تری نسبت به پیکره BNC1 فراهم می‌کند. [۲۸، ۱۹]

- در این رساله، اثر متقابل معیارهای مبتنی بر شبکه واژگان بررسی شده است.

- روش ایزلام و اینکپِن، فقط بسامد واژه (TF) را مورد استفاده قرار داده است ولی از اطلاعات میزان تکرار هر واژه در چند کتاب (IDF) استفاده نکرده است که در روش پیشنهادی محاسبه شده است [۵۹].

- یک واژه ممکن است به صورت اسم، صفت یا قید در جمله بیاید و تمامی نقش‌واژه‌ها در ارزیابی شباهت مهم هستند.

- محاسبه پیچیدگی متن[1] باعث بهبود نتایج همبستگی شده است.

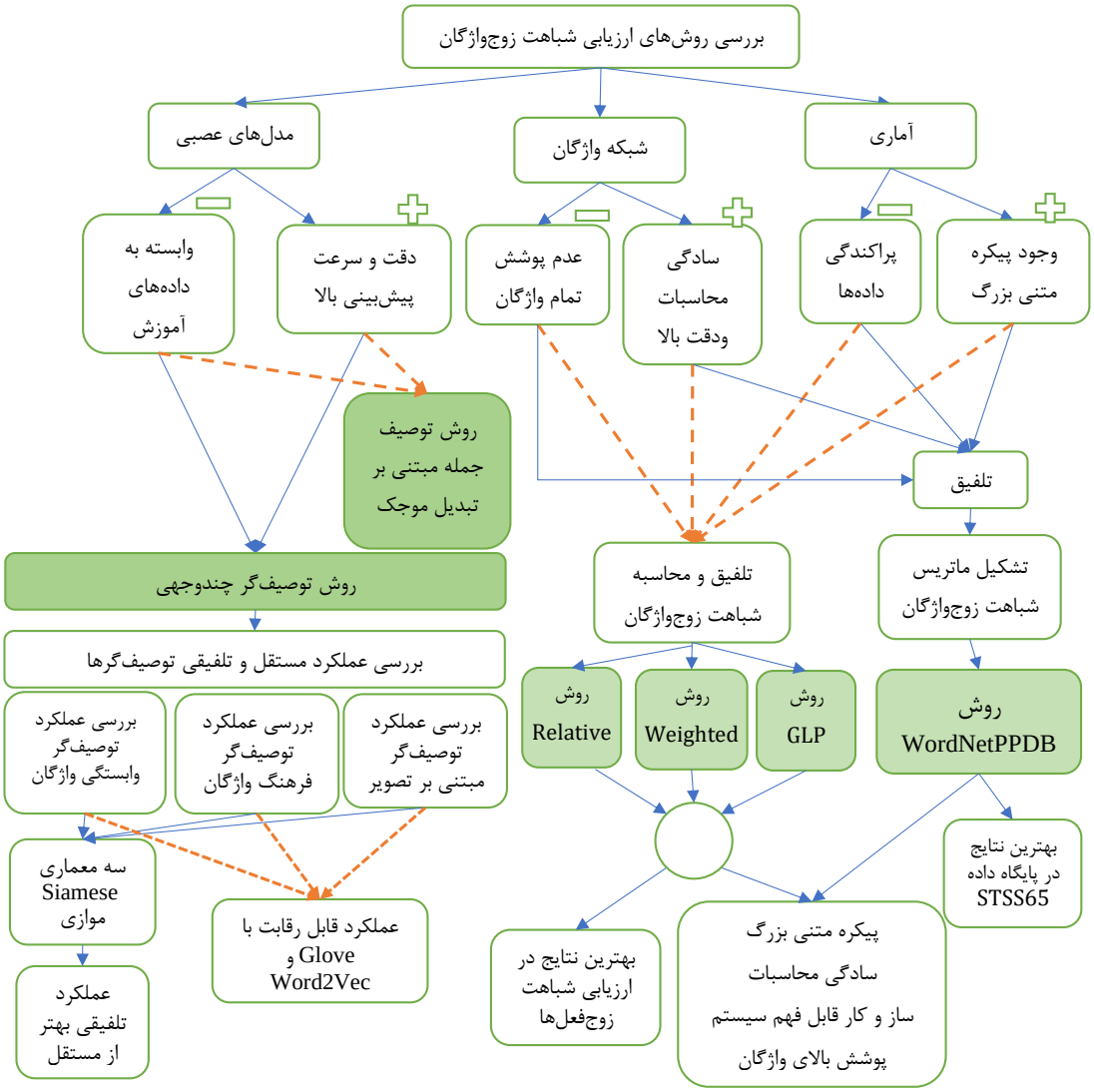
- در این پژوهش، با استفاده از اطلاعات پایگاه داده واژگان PPDB بردار ویژگی برای هر جمله تشکیل شده است سپس با معیارهای مبتنی بر شبکه‌ی واژگان WordNet تلفیق شده است. نتیجه این روش پیشنهادی بهبود در بیشتر آزمون‌های ارزیابی بود.

- عملکرد توصیف‌گرهای مبتنی بر ساختار وابستگی اجزای جمله، توصیف‌گر مبتنی بر اطلاعات

[1] Perplexity

تصاویر و توصیف‌گر مبتنی بر واژگان تعریف در دیکشنری، در ارزیابی شباهت زوج‌جمله‌ها بررسی شده است.

در شکل ۳-۱ فرآیند پژوهش و پیشنهادات پژوهشی آورده شده است. همانطور که در شکل مشخص است هدف اصلی پژوهش تلفیق روش‌های آماری با روش‌های مبتنی بر شبکه واژگان به منظور رفع معایب هرکدام است. در ادامه پژوهش، روش‌های پیشین مبتنی بر مدل‌های عصبی که عمده پژوهش‌های اخیر را تشکیل می‌دهند نیز بررسی شده است. هرکدام از توصیف‌گرها با رویکرد متفاوتی توصیف‌واژگان را می‌سازند. در فصل آینده، عملکرد این روش‌ها بررسی شده است.



تصویر ۳-۱: فرآیند کلی طراحی سیستم‌های پیشنهادی

۳-۱ اطلاعات هم‌رخدادی درجه دوم

روش بعدی که به آن می‌پردازیم اطلاعات هم‌رخدادی نقطه‌به‌نقطه درجه دوم [۲] نام دارد [۶۲]. اگر هدف را محاسبه شباهت بین زوج‌واژه (W_1, W_2) در نظر بگیریم و اگر که تعداد واژه‌های همسایگی از هر طرف را α و اگر بسامد رخداد هر واژه در پیکره را $f^t(t_i)$ در نظر بگیریم که در آن t_i بسامد رخداد واژه i در همسایگی W است، آنگاه فرمول محاسبه تابع اطلاعات نظیر به نظیر برای هم‌رخدادی دو واژه به صورت زیر است:

$$f^{pmi}(t_i, W) = log_2 \frac{f^b(t_i, W) \times m}{f^t(t_i) f^t(W)} \qquad (۱-۳)$$

در این فرمول، نماد f^b بسامد رخداد زوج‌واژه، f^t بسامد رخداد تک‌واژه و f^{pmi} بسامد هم‌رخدادی متقابل نقطه به نقطه است. در صورت کسر $f^b(t_i, W)$ بسامد هم‌رخدادی واژه همسایه t_i با واژه هدف W (در پیکره متن) و m تعداد کل واژگان موجود در پیکره است.

حال فرض بگیرید برای واژه هدف W_1 و واژگان همسایه آن با عنوان گروه Xset فرمول ۳-۱ را محاسبه کردیم. همین عملیات محاسبه برای واژه W_2 و همسایگان آن انجام می‌شود. سپس خروجی‌های تابع فرمول ۳-۱ را به صورت نزولی مرتب کرده و تعداد β_1 نتایج را ذخیره می‌کنیم یعنی طبق فرمول‌های ۳-۲ و ۳-۳ محاسبات انجام می‌شود. همین کار را برای واژه دوم هم انجام می‌دهیم.

همانطور که در فرمول‌های ۳-۲ و ۳-۳ مشخص شده است، واژگان t واژگانی را در بر می‌گیرد که همزمان با W_1 و W_2 هم‌رخداد هستند. در این دو فرمول بسامد هم‌رخدادی آن‌ها به ترتیب نزولی مرتب‌سازی می‌شود. بدین صورت واژگان مشترک مهم در محاسبات مورد استفاده قرار می‌گیرند.

[2] SOC-PMI

$$Xset = Xset_i \; where \; i = 1, 2, ..., \beta_1$$
$$and \; f^{pmi}(t_1, W_1) \geq f^{pmi}(t_2, W_1) \geq f^{pmi}(t_{\beta 2}, W_1) \quad (۲-۳)$$

$$Yset = Yset_i \; where \; i = 1, 2, ..., \beta_2$$
$$and \; f^{pmi}(t_1, W_2) \geq f^{pmi}(t_2, W_2) \geq f^{pmi}(t_{\beta 2}, W_2) \quad (۳-۳)$$

در نتیجه ممقادیر مهم‌تر بسامد نقطه به نقطه واژگان که در مرحله قبل استخراج شد توسط توابع مجموع بتا در فرمول ۳-۴ جمع می‌شود.

$$f^{\beta}(W_1) = \sum_{i=1}^{\beta_1} (f^{pmi}(t_i, W_2))^{\gamma}$$
$$f^{\beta}(W_2) = \sum_{i=1}^{\beta_2} (f^{pmi}(t_i, W_1))^{\gamma} \quad (۴-۳)$$

مقدار $\gamma = 3$ توسط مقاله اصلی پیشنهاد شده است [۶۲]. مقدار نهایی شباهت طبق فرمول ۳-۵ محاسبه می‌شود.

$$sim_{soc-pmi}(W_1, W_2) = \frac{f^{\beta}(W_1)}{\beta_1} + \frac{f^{\beta}(W_2)}{\beta_2} \quad (۵-۳)$$

مقدار بهینه β را می‌توان با فرمول ۳-۶ محاسبه کرد. در این فرمول، $\delta = 6.5$ مقداری ثابت است که مقدار بهینه آن از پژوهش ایزلام و اینکپن گرفته شده است [۶۱]. استفاده از مقادیر کم β_i باعث حذف واژه‌های مهم مشترک می‌شود در حالی که مقدار زیاد آن باعث استفاده از بسامد رخداد واژگان نه چندان مهم می‌شود.

جدول ۳-۱: پیکره متن نمونه

شناسه جمله	جمله
۱	pursuit accident claim car driver exclude
۲	soak motorist company car driver risky
۳	company car driver tend travel farther
۴	job engineer disappear fall mechanical engineer car industry worst affect
۵	sign recession car industry
۶	brightest engineer moment car industry
۷	yugoslavia benefit direct investment automobile industry
۸	acreage expand emergence automobile industry
۹	automobile industry among hardest hit recession
۱۰	automobile industry largely male force
۱۱	component supplier automobile industry expand
۱۲	client industry manufacturer component automobile industry

$$\beta_i = (log(f^t(W_i)))^2 \frac{log_2(n)}{\delta},\ where\ i = 1, 2 \qquad (۶-۳)$$

حل مثال شباهت زوج‌واژه با روش SOC-PMI :

فرض می‌کنیم می‌خواهیم مقدار شباهت بین $W_1 = car$ و $W_2 = automobile$ را محاسبه کنیم. متن مورد نظر ما در جدول ۳-۱ شامل ۱۲ جمله است. بعد از عملیات پیش‌پردازش که شامل حذف ایست‌واژه‌ها و ریشه‌یابی واژگان می‌شود ۷۰ واژه خواهیم داشت که ۴۳ عدد از آن‌ها یکتا خواهند بود. واژگان یکتا و بسامد رخداد آن‌ها در جدول ۳-۲ نمایش داده شده است. در این جدول واژه‌های لیست شده، هم‌رخداد هستند و بسامدهای محاسبه شده مبتنی بر هم‌رخدادی هستند. برای این پیکره متن کوچک مقادیر $\delta = 6.5$ و $\gamma = 3$ استفاده شده است.

حال مقدار نهایی شباهت را به صورت زیر محاسبه می‌کنیم.

جدول ۳-۲: بسامد هم‌رخدادی واژگان هدف با واژگان همسایگی - روش SOC-PMI

شناسه	$x(t_i)$	$f^b(t_i, w_1)$	$f^{pmi}(t_i, w_1)$	شناسه	$Y(t_i)$	$f^b(t_i, w_2)$	$f^{pmi}(t_i, w_2)$
۱	motorist	۱	۳/۵۴۴	۱	emergence	۱	۳/۵۴۴
۲	disappear	۱	۳/۵۴۴	۲	direct	۱	۳/۵۴۴
۳	worst	۱	۳/۵۴۴	۳	acreage	۱	۳/۵۴۴
۴	pursuit	۱	۳/۵۴۴	۴	hit	۱	۳/۵۴۴
۵	soak	۱	۳/۵۴۴	۵	largely	۱	۳/۵۴۴
۶	travel	۱	۳/۵۴۴	۶	yugoslavia	۱	۳/۵۴۴
۷	brightest	۱	۳/۵۴۴	۷	supplier	۱	۳/۵۴۴
۸	fall	۱	۳/۵۴۴	۸	benefit	۱	۳/۵۴۴
۹	risky	۱	۳/۵۴۴	۹	male	۱	۳/۵۴۴
۱۰	company	۲	۳/۵۴۴	۱۰	investment	۱	۳/۵۴۴
۱۱	sign	۱	۳/۵۴۴	۱۱	among	۱	۳/۵۴۴
۱۲	farther	۱	۳/۵۴۴	۱۲	force	۱	۳/۵۴۴
۱۳	accident	۱	۳/۵۴۴	۱۳	client	۱	۳/۵۴۴
۱۴	affect	۱	۳/۵۴۴	۱۴	hardest	۱	۳/۵۴۴
۱۵	mechanical	۱	۳/۵۴۴	۱۵	component	۲	۳/۵۴۴
۱۶	tend	۱	۳/۵۴۴	۱۶	manufacturer	۱	۳/۵۴۴
۱۷	claim	۱	۳/۵۴۴	۱۷	expand	۲	۳/۵۴۴
۱۸	engineer	۳	۳/۵۴۴	۱۸	industry	۷	۳/۰۲۹
۱۹	moment	۱	۳/۵۴۴	۱۹	recession	۱	۲/۵۴۴
۲۰	driver	۳	۳/۵۴۴				
۲۱	exclude	۱	۳/۵۴۴				
۲۲	recession	۱	۲/۵۴۴				
۲۳	industry	۳	۱/۸۰۷				

$$f^{\beta}(W_1) = (f^{pmi}(\text{"}recession\text{"}, W_2))^{\gamma} + (f^{pmi}(\text{"}industry\text{"}, W_2))^{\gamma} = (2/544)^3 + (3/029)^3 = 44/255$$

$$f^{\beta}(W_2) = (f^{pmi}(\text{"}recession\text{"}, W_1))^{\gamma} + (f^{pmi}(\text{"}industry\text{"}, W_1))^{\gamma} = (2/544)^3 + (1/807)^3 = 22/364$$

بنابراین:

$$sim_{socpmi}(W_1, W_2) = \frac{f^{\beta}(W_1)}{\beta_1} + \frac{f^{\beta}(W_2)}{\beta_2} = \frac{44/255}{23} + \frac{22/364}{19} = 3/101$$

سپس برای تعمیم این شیوه محاسبه شباهت زوج‌واژگان به زوج‌جمله‌ها مراحل خاصی باید طی شود.

برای مثال در محاسبه شباهت بین دو جمله:

- A cemetery is a place where dead people's bodies or their ashes are buried.
- A graveyard is an area of land, sometimes near a church, where dead people are buried.

بعد حذف ایست‌واژه‌ها و ریشه‌یابی واژگان باقی‌مانده، ماتریس زیر از محاسبه شباهت زوج‌واژگان بین دو جمله توسط روش SOC-PMI تشکیل می‌شود:

church	near	sometime	land	area	graveyard	
۰/۸۵۶	۰/۵۴۲	۰/۱۹۵	۰/۳۹	۰	۰/۹۸۶	cemetery
۰	۰	۰/۱۴۹	۰/۲۷۶	۰/۴۱۳	۰	place
۰/۰۸۸	۰/۰۶۳	۰/۱۲۲	۰/۳۶۳	۰	۰/۴۶۵	body
۰/۲۱۱	۰/۳۹۵	۰/۲۳۸	۰/۲۱۳	۰	۰/۷۹۶	ash

مهم‌ترین مقادیر شباهت از مراحل زیر استخراج می‌شود:

- ابتدا بزرگ‌ترین عنصر ماتریس پیدا می‌شود و به لیست اضافه می‌شود.

- ردیف و ستون متناظر آن حذف می‌شوند.

- همین‌کار با مقادیر باقی‌مانده ماتریس به صورت زیر ادامه می‌یابد تا زمانی که ماتریس خالی شود.

church	near	sometime	land	area	
۰	۰	۰/۱۴۹	۰/۲۷۶	۰/۴۱۳	place
۰/۰۸۸	۰/۰۶۳	۰/۱۲۲	۰/۳۶۳	۰	body
۰/۲۱۱	۰/۳۹۵	۰/۲۳۸	۰/۲۱۳	۰	ash

church	near	sometime	land	
۰/۰۸۸	۰/۰۶۳	۰/۱۲۲	۰/۳۶۳	body
۰/۲۱۱	۰/۳۹۵	۰/۲۳۸	۰/۲۱۳	ash

church	sometime	land	
۰/۰۸۸	۰/۱۲۲	۰/۳۶۳	body

در نتیجه لیست نهایی p حاوی مقادیر زیر خواهد بود:

$$p = \{0/986, 0/413, 0/363, 0/395\}$$

فرمول نهایی محاسبه شباهت ۳-۷ است. در این فرمول n_1 تعداد واژه‌های جمله‌ی اول و n_2 تعداد واژه‌های جمله‌ی دوم است.

$$sim_{global}(s_1, s_2) = \frac{\alpha \sum_{i=1}^{|p|} p_i \times (n_1 + n_2)}{2 n_1 n_2} \quad (۳-۷)$$

برای مثال، در ارتباط با دو جمله بالا، مقادیر متغیرها به صورت $n_1 = 7, n_2 = 9$ و $\alpha = 3$ است؛ بنابراین فرمول ۳-۷ به صورت زیر محاسبه می‌شود:

$$S = \frac{3(0/986 + 0/413 + 0/363 + 0/395)(7 + 9)}{2 \times 7 \times 9}$$

در یک تغییر نسبت به الگوریتم ایزلام و اینکپِن، تمامی بسامدهای رخداد به صورت نسبی با فرمول ۳-۸ و فرمول ۳-۹ جایگزین می‌شوند. بدین ترتیب برای اشاره به این روش نام این الگوریتم را اطلاعات متقابل هم‌رخدادی نسبی درجه‌ی دو (RSOC-PMI)[3] می‌گذاریم. نسخه‌ای دیگر نیز طراحی شده است تا به واژگان همسایگی وزن بیشتری بدهد بدین ترتیب که در محاسبه بسامد وزن‌های مشابهی مانند $\{1, 2, 3, 4, 5, w, 5, 4, 3, 2, 1\}$ می‌گیرند، نام این نسخه را WSOC-PMI می‌گذاریم.

در فرمول ۳-۸ نماد $booksCount$ تعداد کتاب‌هایی است که در آن‌ها واژه t_i به کار رفته است و در فرمول ۳-۹ نماد تعداد کتاب‌هایی است که زوج‌واژه (t_i, W) کنار هم نوشته شده است.

$$f^t(t_i) = \frac{f^t(t_i)}{booksCount} \tag{۸-۳}$$

$$f^b(t_i, W) = \frac{f^b(t_i, W)}{booksCount} \tag{۹-۳}$$

[3] Relative soc-pmi

۳-۲ تطابق رشته‌ها بین واژگان

روش بعدی مورد استفاده، محاسبه نرخ یکسان بودن نویسه‌های استفاده شده بین واژگان دو جمله است که خود از سه فرمول محاسبه می‌شود. در این روش، طولانی‌ترین زیردنباله یکسان با نماد LCSS[۴] بین واژگان محاسبه می‌شود و سپس نسبت به طول کل دو واژه مورد مقایسه، نرمال می‌شود (NLCSS[۵] فرمول ۳-۱۰).

$$v_1 = NLCSS(W_1, W_2) = \frac{len(LCSS(W_1, W_2))^2}{len(W_1) \times len(W_2)} \qquad (۱۰-۳)$$

در فرمول ۳-۱۰ نماد len به معنای تعداد نویسه‌های رشته است. در فرمول ۳-۱۱، طولانی‌ترین زیربخش واژه کوتاه‌تر (از نظر تعداد نویسه‌ها) که در واژه طولانی‌تر وجود دارد را محاسبه می‌کند، به شرط آنکه این محاسبه از نویسه اول شروع شده باشد NMCLCSS[۶]. فرمول ۳-۱۲ همین محاسبه را انجام می‌دهد با این تفاوت که طولانی‌ترین زیربخش می‌تواند از هر نویسه‌ای آغاز شود.

$$v_2 = NMCLCSS_1(W_1, W_2) = \frac{len(MCLCSS_1(W_1, W_2))^2}{len(W_1) \times len(W_2)} \qquad (۱۱-۳)$$

$$v_3 = NMCLCSS_n(W_1, W_2) = \frac{len(MCLCSS_n(W_1, W_2))^2}{len(W_1) \times len(W_2)} \qquad (۱۲-۳)$$

عدد نهایی توسط این سه فرمول یاد شده با عنوان شباهت مبتنی بر رشته، توسط فرمول ۳-۱۳

[4] Longest common subsequence

[5] Normalized LCSS

[6] Normalized maximal consecutive LCSS

محاسبه می‌شود. پژوهش [۵۹] مقدار $\alpha = 0/33$ پیشنهاد داده است.

$$Sim_{string}(W_1, W_2) = \alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3 \quad (۱۳-۳)$$

برای مثال برای زوج‌واژه (place, land) به مقادیر زیر می‌رسیم:

$$v_1 = \frac{2^2}{5 \times 4} = 0/2$$

$$v_2 = 0$$

$$v_3 = \frac{2^2}{5 \times 4} = 0/2$$

$$Sim_{string} = 0/33 \times 0/2 + 0/33 \times 0 + 0/33 \times 0/2 = 0/132$$

به‌طور کلی ماتریس محاسبات به روش تطابق نویسه‌ها به صورت زیر است:

church	near	sometime	land	area	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰	۰/۰۲۱	۰/۰۲۳	cemetery
۰/۰۲۲	۰/۰۳۳	۰/۰۱۷	۰/۱۳۲	۰/۰۸۳	۰/۰۳۷	place
۰	۰	۰/۰۲۱	۰/۰۴۱	۰	۰/۰۱۸	body
۰/۰۳۷	۰/۰۵۵	۰/۰۲۸	۰/۰۵۵	۰/۰۸۳	۰/۰۲۴	ash

church	near	sometime	area	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰/۰۲۱	۰/۰۲۳	cemetery
۰	۰	۰/۰۲۱	۰	۰/۰۱۸	body
۰/۰۳۷	۰/۰۵۵	۰/۰۲۸	۰/۰۸۳	۰/۰۲۴	ash

جدول ۳-۳: مقایسه درستی تشخیص سه روش آماری

LSA	PMI-IR	SOC-PMI
۶۴/۳۷٪	۷۳/۷۵٪	۷۶/۲۵٪

church	near	sometime	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰/۰۲۳	cemetery
۰	۰	۰/۰۲۱	۰/۰۱۸	body

church	sometime	graveyard	
۰	۰/۰۲۱	۰/۰۱۸	body

در نتیجه لیست p حاوی مقادیر زیر خواهد بود:

$$p = \{0/132, 0/083, 0/129, 0/021\}$$

و نتیجه شباهت به صورت زیر از فرمول۳-۷ محاسبه می‌شود:

$$S = \frac{3(0/132 + 0/083 + 0/129 + 0/021)(7+9)}{2 \times 7 \times 9} = 0/139$$

جدول ۳-۳ نتایج ارزیابی سه روش یاد شده را نمایش می‌دهد که توسط پایگاه‌های داده‌ای که در ادامه لیست شده، آزموده شده‌اند:

- ۸۰ پرسش آزمون انتخاب مترادف تافل.
- ۵۰ پرسش امتحان انگلیسی به عنوان زبان دوم (ESL).
- رسش انتخاب جفت نام مرتبط با (نقش‌واژه) اسم نوشته شده از پایگاه داده Miller [۸۸].
- ۶۵ پرسش انتخاب جفت نام مرتبط از پایگاه داده Rubbenstein [۱۱۰].

۳-۳ کمیت قدرت کل

این معیار ارزیابی[7] ربط دو واژه فرمولی مستقیم برای محاسبه ارتباط بسامد هم‌رخدادی دو عبارت یا واژه است. ما این معیار را به دو روش برای محاسبه روابط بسامد کل و بسامد صفحه استفاده کردیم [۱۰۷].

- بسامد کل: تعداد کل تکرار یک واژه یا عبارت در کل اسناد.
- بسامد صفحه: تعداد صفحه‌هایی که عبارت در آن تکرار شده است صرف‌نظر از آنکه چه تعداد در یک سند تکرار شده باشد.

فرمول ۳-۱۴ نحوه محاسبه این معیار است. دقت بفرمایید که دو بسامد تکرار نزدیک‌تر باعث می‌شوند این معیار عدد کوچک‌تری تولید کند. $f^t(W_1)$ بسامد تکرار واژه W_1 است و $f^t(W_2)$ بسامد تکرار واژه W_2 است.

$$sim_{qts}(W_1, W_2) = (f^t(W_1) + f^t(W_2)) \times \frac{min(f^t(W_1), f^t(W_2))}{max(f^t(W_1), f^t(W_2))} \qquad (۱۴-۳)$$

۴-۳ مدل سه‌نویسه‌ای گوگل

این مدل[8] ارتباط بین زوج‌واژگان را با فرمول ۳-۱۵ می‌سنجد. برای این رویکرد فقط مقادیر بسامد کل واژگان را استفاده کردیم. برای هر واژه فهرستی از ۱۰۰ سه‌نویسه جستجو می‌شود. این مدل را به چهارنویسه و بیشتر تعمیم دادیم، همچنین از بسامد صفحه واژگان استفاده کردیم اما به دلیل بهبود ندادن نتایج، آن‌ها را حذف کردیم. در فرمول ۳-۱۵، نماد $tf(W_1, t_i, W_2)$ بسامد هم‌رخدادی دو واژه (W_1, W_2) را محاسبه می‌کند در صورتی که با فاصله یک واژه (t_i) از هم جدا باشند. در بخش $tf(W_2, t_i, W_1)$ همین محاسبات دوباره انجام می‌شود با این تفاوت که جای واژه‌ها عوض می‌شود.

[7] Quantified total strength

[8] Google tri-gram model

$$gtm(W_1, W_2) = \frac{1}{2}(\sum_i tf(W_1, t_i, W_2) + \sum_j tf(W_2, t_j, W_1)) \quad (۱۵-۳)$$

بعد این مرحله، فرمول ۳-۱۶ برای نرمال‌سازی نتیجه به بازه صفر و یک استفاده می‌شود [۶۰]. فرمول ۳-۱۷ نتیجه نهایی شباهت را تولید می‌کند. در این فرمول C بسامد رخداد پرتکراترین واژه در پیکره است.

$$f^S(w_1, w_2) = \frac{gtm(W_1, W_2) \times C^2}{f^t(W_1) f^t(W_2) min(f^t(W_1), f^t(W_2))} \quad (۱۶-۳)$$

$$sim_{gtm}(W_1, W_2) = \left\{ \begin{array}{ll} \frac{log(f^S(W_1, W_2))}{denominator} & if\ f^S > 1 \\ \frac{log(1.01)}{denominator} & if\ f^S \leq 1 \\ 0 & if\ f^S = 0 \end{array} \right\} \quad (۱۷-۳)$$

$$denominator = -2 \times log(\frac{min(f^t(W_1), f^t(W_2))}{C})$$

در این فرمول نماد f^S بسامد اسکیپ‌گرام زوج‌واژه است و نماد $denominator$ مخرج کسر است که برای خوانایی بیشتر در خطی مجزا نوشته شده است.

Word2Set ۵-۳

روش «تبدیل واژه به مجموعه[۹]» در گراف شبکه واژگان جستجو می‌کند و برای هر واژه، یک مجموعه شامل واژگان همسایه و کلیدواژه‌ها (از تعریف متنی) تشکیل می‌شود. اجازه دهید تا R_1 مجموعه واژه‌های مرتبط به W_1 و R_2 مجموعه واژه‌های مرتبط به W_2 است. شباهت توسط فرمول ۳-۱۸ محاسبه می‌شود.

$$sim_{word2set} = \frac{|R_1 \cap R_2|}{\sqrt{|R_1| \times |R_2|}} \quad (۱۸-۳)$$

همانطور که در فرمول مشخص است، معیارهای اصلی محاسبه شباهت، شامل تعداد اشتراک واژه‌های مرتبط به هم بین دو واژه است.

۶-۳ معرفی معیار وابستگی فاصله‌ی وزنی مسیر-ووپالمر-لی

این معیار از سه فرمول ۲-۳، ۲-۴ و ۲-۵ مشتق شده است. ایده این معیار آن است که همواره فاصله‌های سنجیده شده خطی نیستند و ممکن است بین معیارهای ذکرشده، روابط غیرخطی وجود داشته باشد که در اینجا با فرمول ۳-۱۹ محاسبه شده است. همچنین در صورتی که یکی از این معیارها برای مقدار شباهت صفر در نظر گرفته شود، خروجی صفر می‌شود و مقادیر بقیه معیارها نیز در نظر گرفته نمی‌شود؛ بنابراین در صورتی دو واژه مقدار شباهت خواهند داشت که هر سه معیار مقدار شباهت به آن اختصاص دهند. شباهت دو جمله می‌تواند با استفاده از ماتریس شباهت زوج‌واژگان محاسبه شود.

$$sim_{wwl} = 3 \times sim_{wpath} \times sim_{wup} \times sim_{li} \quad (۱۹-۳)$$

[9] Word2Set

۳-۷ معرفی معیار وابستگی فاصله‌ی وزنی مسیر-لی

این معیار از دو فرمول ۲-۴ و ۲-۵ مشتق شده است. این دو معیار نسبت به معیار فرمول ۲-۳ به هم شباهت بیشتری دارند زیرا که هر دو کوتاه‌ترین مسیر بین دو مفهوم را در نظر می‌گیرند ولی در فرمول ۲-۳ در نظر گرفته نمی‌شود. فرمول محاسبه این معیار در زیر آورده شده است.

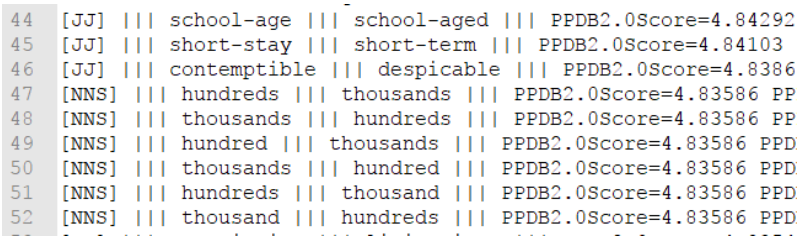
$$sim_{wwl} = 2 \times sim_{wpath} \times sim_{li} \quad (۲۰-۳)$$

۳-۸ شباهت مبتنی بر پایگاه دانش PPDB

این پایگاه داده حاوی لیست بزرگی از واژه‌های قابل جایگزین همراه با نمره شباهت دو واژه است. این نمره بر اساس مقیاس لیکِرت بین صفر تا پنج ارزش‌گذاری می‌شود [۹۹]. نمره نهایی میانگین ارزیابی نمره‌های تخصیص داده شده توسط داوران انسانی است. این مقیاس در زیر توضیح داده شده است، همچنین نمونه‌ای از داده‌های این پایگاه دانش در تصویر ۳-۲ نمایش داده شده است. مقیاس لیکرت:

- کاملاً مخالف (قطعاً بدون هیچ شباهتی)
- مخالف (بدون هیچ شباهت)
- نه موافق و نه مخالف (نه شبیه و نه بدون شباهت)
- موافق (شبیه)
- کاملاً موافق (کاملاً قابل جایگزین)

برای پیاده‌سازی این بخش، از روش معرفی شده توسط کرافت و همکاران استفاده شده است [۲۹]؛ با این تفاوت که به جای معیارهای WordNet از پایگاه دانش PPDB استفاده شده است. لازم به ذکر



تصویر ۳-۲: نمونه‌ای از پایگاه دانش PPDB

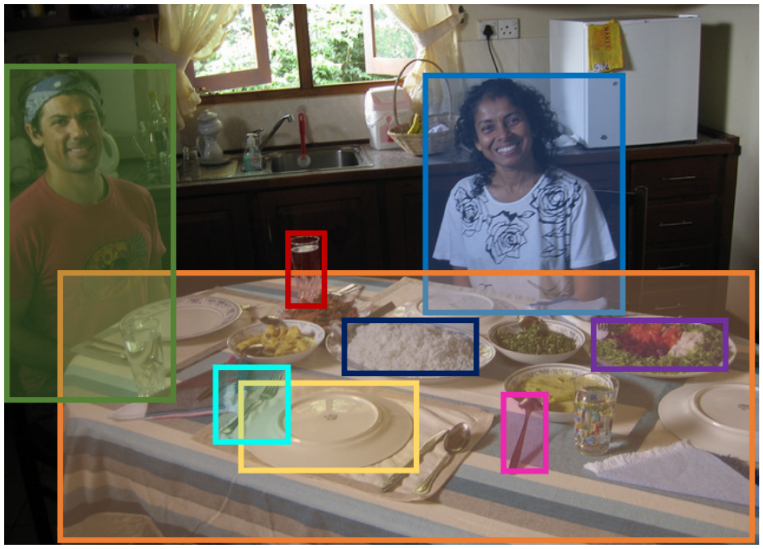
است که تمامی مقدارها در بازه صفر و یک نرمال‌سازی شده است. دلیل این پیاده‌سازی آن است که این روش می‌تواند مشکل کمبود داده‌های کافی برای مقایسه دو جمله را تا حدودی برطرف کند و همچنین نتایج خوبی در محاسبه شباهت داشته باشد. در این روش از دو جمله، یک لیست حاوی واژگان یکتا استخراج می‌شود. سپس تمامی واژه‌های لیست با واژه‌های جمله‌ی اول مقایسه شده و شباهت محاسبه می‌شود، همین فرآیند در مورد جمله‌ی دوم تکرار می‌شود؛ توضیح کامل این روش به عنوان روش پیشین در فصل دو جدول ۲-۴ آورده شده است.

۳-۹ توصیفگر Vico

به جای در نظر گرفتن واژگان همسایه در متن، برای تولید بردار ویژگی واژه از اشیاء نزدیک به هم در تصاویر به عنوان واژگان همسایه استفاده شده است [۴۱]. اما روش تولید بردار ویژگی جدا از این مسئله مانند روش Glove است. به عبارتی همانطور که در تصویر ۳-۳ نیز دیده می‌شود، زوج‌واژه «مرد و زن» می‌توانند هم‌رخدادی بالایی داشته باشند اگر که در تصاویر کنار هم قرار بگیرند. اشیاء روی میز نیز به دلیل هم‌رخدادی بالا توصیف‌گر مشابهی خواهند داشت.

۳-۱۰ توصیفگر SynGCN

پژوهشگران با استفاده از گراف محاسباتی شبکه عصبی پیچشی و اطلاعات تجزیه بردار وابستگی واژگان جمله، توصیفگرهای SynGCN را ساخته‌اند [۱۲۸]. درخت وابستگی توسط ابزار آماده استنفورد [۸۱] تشکیل می‌شود. در این پژوهش به واژه هدف و همسایگانش به عنوان یک گراف به مرکزیت واژه



تصویر ۳-۳: نمونه‌ای از تصویرمرتبط با vico [۴۱]

هدف نگاه می‌شود. ابتدا توصیفگرها با مقادیر تصادفی شروع می‌شوند. هدف الگوریتم بهینه‌ساز پیدا کردن تلفیقی از گره‌های همسایه برای گره مرکزی تعریف شده است. هر گره همسایه (با توجه به درخت وابستگی آن) ممکن است به صورت متغیری به واژه مرکزی اضافه شود و بنابریان هدف کدگذاری شباهت عملکردی تعریف می‌شود. به عبارتی واژگان شبیه قابل جایگزینی با یکدیگر هستند.

۳-۱۱ توصیفگر Dict2Vec

در این پژوهش [۱۲۲]، بحث می‌شود که واژگان همسایه در بافت متن لزوماً توصیف مشابهی با واژه مرکزی ندارند. به عبارتی این واژگان ممکن است از نظر معنایی متفاوت یا بی‌ربط باشند و درنتیجه عملکرد مناسبی نداشته باشند. فرهنگ لغت[۱۰] شامل توصیف واژگان بهتری است زیرا که واژگان مرتبط‌تر در آن به واژگان کلیدی شبیه‌تر هستند تا واژه‌هایی که در همسایگی یک متن نمونه یافت می‌شوند. علاوه بر این، اطلاعات بیشتری نسبت به شبکه واژگان فراهم شده است و برخلاف آن، توصیف‌های واژگان مشخص‌تر و بهتر تعریف می‌شود. هر ورودی کتاب فرهنگ واژگان شامل دو کلید است به نام‌های «کلیدواژه و واژگان توصیف‌کننده». واژگانی که به صورت متقابل در تعریف خود واژه کلیدی را دربرداشته

[10]Dictionary

باشند به آن‌ها زوج‌های قوی و در غیر این صورت زوج‌های ضعیف گفته می‌شود. هدف بهینه‌سازی کاهش فاصله بین توصیفگرهای واژه و زوج آن‌ها است. بدین صورت هر نوع زوج (قوی یا ضعیف) به پارامترهای آموزش سیستم، به صورت متفاوتی اضافه می‌شوند [۱۲۲].

۳-۱۲ استخراج ویژگی‌های محلی از ماتریس شباهت

تأثیر استخراج ویژگی‌های محلی از ماتریس شباهت تمام زوج‌واژگان هم بررسی شده است، بار دیگر ماتریس زیر را در نظر بگیرید. برای محاسبه در این بخش، یک پنجره 3×3 به صورت شناور روی ماتریس می‌گذاریم و میانگین عنصرهای ماتریس را محاسبه کرده و در لیست p قرار می‌دهیم؛ مراحل این کار در زیر نمایش داده شده است.

church	near	sometime	land	area	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰	۰/۰۲۱	۰/۰۲۳	cemetery
۰/۰۲۲	۰/۰۳۳	۰/۰۱۷	۰/۱۳۲	۰/۰۸۳	۰/۰۳۷	place
۰	۰	۰/۰۲۱	۰/۰۴۱	۰	۰/۰۱۸	body
۰/۰۳۷	۰/۰۵۵	۰/۰۲۸	۰/۰۵۵	۰/۰۸۳	۰/۰۲۴	ash

$$p_1 = \frac{0/023 + 0/021 + 0 + 0/037 + 0/083 + 0/132 + 0/018 + 0 + 0/041}{9} = 0/0394$$

church	near	sometime	land	area	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰	۰/۰۲۱	۰/۰۲۳	cemetery
۰/۰۲۲	۰/۰۳۳	۰/۰۱۷	۰/۱۳۲	۰/۰۸۳	۰/۰۳۷	place
۰	۰	۰/۰۲۱	۰/۰۴۱	۰	۰/۰۱۸	body
۰/۰۳۷	۰/۰۵۵	۰/۰۲۸	۰/۰۵۵	۰/۰۸۳	۰/۰۲۴	ash

$$p_2 = \frac{0/021 + 0/083 + 0 + 0 + 0/132 + 0/041 + 0/129 + 0/017 + 0/021}{9} = 0/0493$$

church	near	sometime	land	area	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰	۰/۰۲۱	۰/۰۲۳	cemetery
۰/۰۲۲	۰/۰۳۳	۰/۰۱۷	۰/۱۳۲	۰/۰۸۳	۰/۰۳۷	place
۰	۰	۰/۰۲۱	۰/۰۴۱	۰	۰/۰۱۸	body
۰/۰۳۷	۰/۰۵۵	۰/۰۲۸	۰/۰۵۵	۰/۰۸۳	۰/۰۲۴	ash

$$p_3 = \frac{0 + 0/132 + 0/041 + 0/129 + 0/017 + 0/021 + 0/052 + 0/033 + 0}{9} = 0/0472$$

church	near	sometime	land	area	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰	۰/۰۲۱	۰/۰۲۳	cemetery
۰/۰۲۲	۰/۰۳۳	۰/۰۱۷	۰/۱۳۲	۰/۰۸۳	۰/۰۳۷	place
۰	۰	۰/۰۲۱	۰/۰۴۱	۰	۰/۰۱۸	body
۰/۰۳۷	۰/۰۵۵	۰/۰۲۸	۰/۰۵۵	۰/۰۸۳	۰/۰۲۴	ash

$$p_4 = \frac{0/129 + 0/017 + 0/021 + 0/052 + 0/033 + 0 + 0/041 + 0/02 + 0}{9} = 0/0347$$

church	near	sometime	land	area	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰	۰/۰۲۱	۰/۰۲۳	cemetery
۰/۰۲۲	۰/۰۳۳	۰/۰۱۷	۰/۱۳۲	۰/۰۸۳	۰/۰۳۷	place
۰	۰	۰/۰۲۱	۰/۰۴۱	۰	۰/۰۱۸	body
۰/۰۳۷	۰/۰۵۵	۰/۰۲۸	۰/۰۵۵	۰/۰۸۳	۰/۰۲۴	ash

$$p_5 = \frac{0/037 + 0/018 + 0/024 + 0/083 + 0 + 0/083 + 0/132 + 0/041 + 0/055}{9} = 0/0525$$

church	near	sometime	land	area	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰	۰/۰۲۱	۰/۰۲۳	cemetery
۰/۰۲۲	۰/۰۳۳	۰/۰۱۷	۰/۱۳۲	۰/۰۸۳	۰/۰۳۷	place
۰	۰	۰/۰۲۱	۰/۰۴۱	۰	۰/۰۱۸	body
۰/۰۳۷	۰/۰۵۵	۰/۰۲۸	۰/۰۵۵	۰/۰۸۳	۰/۰۲۴	ash

$$p_6 = \frac{0/083 + 0 + 0/083 + 0/132 + 0/041 + 0/055 + 0/017 + 0/021 + 0/028}{9} = 0/0511$$

church	near	sometime	land	area	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰	۰/۰۲۱	۰/۰۲۳	cemetery
۰/۰۲۲	۰/۰۳۳	۰/۰۱۷	۰/۱۳۲	۰/۰۸۳	۰/۰۳۷	place
۰	۰	۰/۰۲۱	۰/۰۴۱	۰	۰/۰۱۸	body
۰/۰۳۷	۰/۰۵۵	۰/۰۲۸	۰/۰۵۵	۰/۰۸۳	۰/۰۲۴	ash

$$p_7 = \frac{0/132 + 0/041 + 0/055 + 0/017 + 0/021 + 0/028 + 0/033 + 0 + 0/055}{9} = 0/0424$$

church	near	sometime	land	area	graveyard	
۰/۰۴۱	۰/۰۵۲	۰/۱۲۹	۰	۰/۰۲۱	۰/۰۲۳	cemetery
۰/۰۲۲	۰/۰۳۳	۰/۰۱۷	۰/۱۳۲	۰/۰۸۳	۰/۰۳۷	place
۰	۰	۰/۰۲۱	۰/۰۴۱	۰	۰/۰۱۸	body
۰/۰۳۷	۰/۰۵۵	۰/۰۲۸	۰/۰۵۵	۰/۰۸۳	۰/۰۲۴	ash

$$p_8 = \frac{0/017 + 0/021 + 0/028 + 0/033 + 0 + 0/055 + 0/022 + 0 + 0/037}{9} = 0/0236$$

حال با استفاده از فرمول ۳-۷ نتیجه شباهت را محاسبه می‌کنیم:

$$sim_{global}(s_1, s_2) = \frac{3 \times (0/0394 + 0/0493 + 0/0472 + 0/0347 + 0/0525 + 0/0511 + 0/0424 + 0/0236) \times (7 + 9)}{2 \times 7 \times 9}$$

۳-۱۳ روش‌های پیشنهادی مبتنی بر اطلاعات آماری

این روش‌های مبتنی بر یک روش آماری هستند که از همسایگان مشترک دو واژه برای محاسبه شباهت بین آن‌ها استفاده می‌کند.

۳-۱۳-۱ محاسبه شباهت تلفیقی زوج‌واژگان

یک رگرسیون برای انتخاب ضرایب سه ویژگی روش‌های:

- روش $sim_{word2set}$ از بخش ۳-۵
- محاسبه sim_{qts} برای بسامد رخداد کل از بخش ۳-۳
- محاسبه sim_{qts} برای بسامد رخداد تعداد صفحات تکرار شده از بخش ۳-۳

طراحی می‌شود. آموزش این رگرسیون، روی پایگاه داده پیشنهادی $W1500$ پژوهش خیمنز و همکاران انجام می‌شود [۶۶]. این پایگاه داده شامل: ۵۰۰ واژه مترادف با مقدار ربط تنظیم شده یک، ۵۰۰ واژه متضاد با مقدار ربط ۰/۲ و ۵۰۰ واژه نامرتبط با مقدار ربط صفر است. رویکرد ما، مبتنی بر مهندسی ویژگی است. در این رویکرد، برای محاسبه شباهت یک زوج‌واژه، یک بردار حاوی سه ویژگی محاسبه می‌کنیم. در این رویکرد، هم روابط بسامد خود واژه (بسامد مستقیم) با روش «کمیت قدرت کل» و هم «اطلاعات واژگان هم‌رخداد» را با روش Word2Set در نظر می‌گیریم. سپس این مقادیر به یک رگرسیون ارسال می‌شوند. مدل رگرسیون یک «فرآیند گاوسین[۱۱] » با استفاده از «تابع پایه شعاعی[۱۲]» است. اضافه کردن نویز و اعتبارسنجی متقابل[۱۳]، رویکرد ما برای پیشگیری از بیش‌برازش[۱۴] است. مدل رگرسیون مورد استفاده GPR یک مدل رگرسیون بیزیَن چند متغیره و غیرپارامتری[۱۵] است. دلیل انتخاب هرکدام از ویژگی‌ها به شرح زیر است:

[11] Gaussian process (GPR)
[12] Radial basis function
[13] Cross validation
[14] Overfitting
[15] Non-parametric

۱. روش Word2Set نوین‌ترین و بهترین نتایج را در زمینه ارزیابی شباهت واژگان کسب کرده است. این روش بسیار شهودی است چراکه واژگان مشترک در بسط تعریف دو واژه برای محاسبه شباهت استفاده می‌شوند.

۲. فرمول آماری مورد استفاده، برای «بسامد کل[۱۶]» و «بسامد رخداد در سند[۱۷]» به صورت مستقل محاسبه می‌شود (دو ویژگی). در نتیجه، در صورت نبود واژه‌ای در شبکه واژگان، این دو مقدار می‌توانند ارزیابی خوبی را «به عنوان جایگزین» ارائه دهند، ضمن آنکه اطلاعات بسامد رخداد از «پیکره‌های متنی بزرگ گوگل» محاسبه می‌شود.

۳-۱۳-۲ روش پیشنهادی اول: مبتنی بر واژگان همسایگی وزن‌دهی شده (Weighted)

این روش پیشنهادی خود از سه روش زیر تشکیل شده است:

۱. تطابق رشته‌ها بین واژگان در بخش ۳-۲

۲. اطلاعات متقابل هم‌رخدادی وزنی درجه‌ی دو (WSOC-PMI) از بخش ۳-۱

۳. معیار مبتنی بر شبکه‌ی واژگان لی مرور شده در بخش ۲-۲-۲

برای هرکدام از این معیارها ماتریس شباهت زوج‌واژگان با نماد M محاسبه می‌شود. در نتیجه برای سه روش یاد شده سه ماتریس M خواهیم داشت. سپس با فرمول ۳-۲۱ میانگین سه ماتریس در یک ماتریس خلاصه می‌شود، به عبارتی هر عنصر ماتریس M نهایی میانگین مقدار عنصر در سه ماتریس است.

$$M(i,j) = \frac{1}{n}\sum_{k=1}^{n} M_k(i,j) \qquad (۲۱-۳)$$

[16] Term frequency

[17] Inverse document frequency

که در آن n تعداد ماتریس‌های مورد استفاده است که در اینجا سه است. در انتها با فرمول ۳-۷ شباهت محاسبه می‌شود.

۳-۱۳-۳ روش پیشنهادی دوم: هم‌رخدادی واژه‌ها و شباهت مفهوم‌ها (Relative)

این روش پیشنهادی از دو روش زیر تشکیل شده است:

- اطلاعات متقابل هم‌رخدادی نسبی درجه‌ی دو (RSOC-PMI) در بخش ۳-۱
- معیار وابستگی فاصله‌ی وزنی مسیر-ووپالمر-لی در بخش ۳-۶

برای هرکدام از این معیارها ماتریس M محاسبه می‌شود اما این‌بار با استفاده از داده‌های آموزش وزن متفاوتی برای هرکدام از معیارها در محاسبه ماتریس کل در نظر گرفته می‌شود:

$$M(i,j) = \frac{1}{n}\sum_{k=1}^{n} \alpha_k M_k(i,j) \qquad (۲۲-۳)$$

که در این فرمول، α مقداری بین صفر تا یک می‌گیرد و جمع تمام ضرایب باید یک شود. البته در آزمایش‌ها مقدار بهینه برای روش هم‌رخداد $\alpha_1 = 0/75$ و در نتیجه، برای روش معیارهای تلفیقی $\alpha_2 = 0/25$ به‌دست آمده است که قابل‌تعمیم به پایگاه داده‌های مختلف است. در انتها با فرمول ۳-۷ شباهت محاسبه می‌شود.

۳-۱۳-۴ روش پیشنهادی سوم: مبتنی بر ویژگی‌های عمومی و محلی (GLP)

در روش GLP[18] بزرگ‌ترین عناصر ماتریس شباهت زوج‌واژگان به عنوان ویژگی‌های عمومی استخراج می‌شود و ویژگی‌های توضیح داده شده در بخش ۳-۱۲ که در آن از پنجره شناور استفاده می‌شود به عنوان ویژگی‌های محلی محاسبه می‌شوند. هدف از استفاده از ویژگی‌های محلی جلوگیری از حذف

[18] Global and local features

اطلاعات شباهت‌های واژگان دور از زوج‌واژه مورد مقایسه است. این اطلاعات ممکن است مهم باشد اما در رویکرد عمومی به دلیل حذف یک ردیف و ستون کامل ممکن است از دست برود.

این روش پیشنهادی خود از سه روش زیر تشکیل شده است:

۱. تطابق رشته‌ها بین واژگان در بخش ۳-۲

۲. معیار وابستگی فاصله‌ی وزنی مسیر-ووپالمر-لی در بخش ۳-۶

۳. پیچیدگی[19] جمله

شباهت مبتنی بر پیچیدگی جمله روی تک‌نویسه‌ها[20] طبق فرمول ۳-۲۵ محاسبه شده است [۱۳۴، ۲۰]. لیست تک‌نویسه‌ها را می‌توان بعد از به هم پیوستن دو جمله با نماد s به دست‌آورد، به عبارتی نویسه‌های دو جمله از هم جدا می‌شوند و نویسه‌های یکتا در لیست ذخیره می‌شوند.

$$entropy(s) = H(s) = -\sum_{x} p(x) log_2 p(x) \qquad (۲۳-۳)$$

$$perplexity(s) = pp(s) = 2^{H(s)} \qquad (۲۴-۳)$$

$$sim_{perplexity}(s_1, s_2) = 1 - \frac{perplexity(s)}{numberOfUnique1grams} \qquad (۲۵-۳)$$

در فرمول ۳-۲۳، نماد $p(x)$ احتمال رخداد هر تک‌نویسه در دوجمله است. برای روش اول و دوم

[19] Perplexity

[20] Unigrams

ماتریس شباهت زوج‌واژگان M را به‌دست می‌آوریم و بعد از استخراج لیست p از ویژگی‌های عمومی و محلی روی ماتریس (دو لیست p_1 و p_2)، مقدار نهایی شباهت در دو حالت عمومی و محلی شباهت زوج‌جمله از ماتریس را به صورت فرمول ۳-۷ محاسبه می‌کنیم سپس از آن‌ها طبق فرمول ۳-۲۶ میانگین می‌گیریم:

$$S_{total} = \frac{sim_{global} + sim_{local} + sim_{perplexity}}{3} \quad (۲۶-۳)$$

در اینجا sim_{global} نماد شباهت زوج‌جمله مبتنی بر ویژگی‌های عمومی و sim_{local} نماد شباهت مبتنی بر ویژگی‌های محلی است.

۳-۱۳-۵ روش پیشنهادی چهارم: مبتنی بر WordNet و PPDB (WordnetPPDB)

همان‌طور که از عنوان مشخص است این روش از دو بخش تشکیل می‌شود:

۱. مبتنی بر معیار شبکه‌ی واژگان: روش مسیر وزن‌دهی در بخش ۲-۲-۳. برای این ماتریس شباهت زوج‌واژگان تشکیل شده است و ویژگی‌های عمومی محاسبه شده است.

۲. مبتنی بر شبکه‌ی واژگان قابل جایگزین PPDB: پیاده‌سازی این بخش کاملاً مشابه با روش کرافت و همکاران است [۲۹] با این تفاوت که معیار شباهت تغییر کرده است.

نتیجه‌ی نهایی شباهت دو جمله، میانگین ارزیابی شباهت هر دو معیار است.

۳-۱۴ مدل‌های عصبی

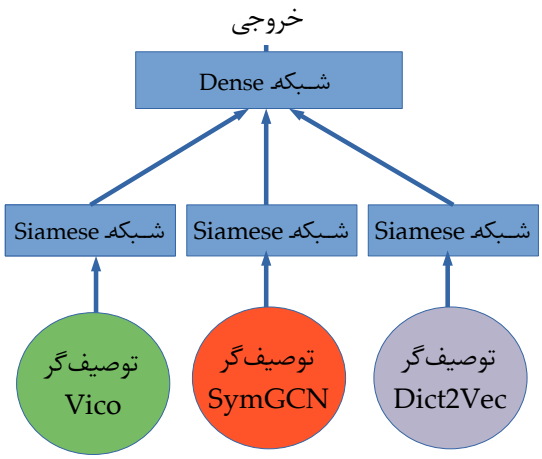
در این روش‌های از معماری‌های شبکه عصبی استفاده شده است. در ادامه توضیحات این شبکه‌ها و معماری‌های مورد استفاده آمده است.

۳-۱۴-۱ شباهت زوج‌جمله - توصیف چندوجهی

استفاده انحصاری از یک نوع توصیفگر تصویر یکنواختی از توصیف واژگان برای تشکیل بردار توصیف جمله ارائه می‌دهد. به عبارتی هر روش توصیفگر یک دید انحصاری از واژگان را داراست برای مثال Word2Vec یا Glove اطلاعات هم‌رخدادی واژگان را به عنوان توصیف واژه می‌سازند. اما برای حل این مشکل و ارائه یک توصیف متنوع و چندوجهی از واژگان ما از سه نوع روش متفاوت برای توصیف واژگان استفاده کردیم. بردار تولید شده توسط این سه روش با یکدیگر توصیف کامل‌تری نسبت به حالت مستقل، ارائه می‌دهند. این سه روش شامل موارد زیر است:

- توصیف‌گر SynGCN از بخش ۳-۱۰
- توصیف‌گر Vico از بخش ۳-۹
- توصیف‌گر Dict2Vec از بخش ۳-۱۱

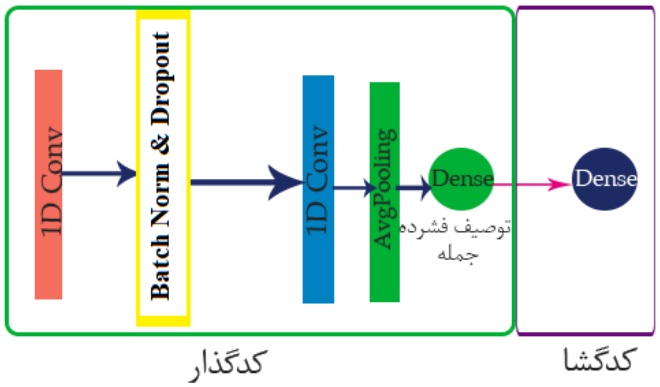
معماری مورد استفاده برای آموزش شبکه در تصویر ۳-۴ نمایش داده شده است و بر پایه آنچه که در معماری Siamese تعریف شده طراحی شده است. در تغییری دیگر عملکرد GRU، BiGRU و BiLSTMها آزموده شد و بهترین عملکرد را LSTM دوطرفه داشت.



تصویر ۳-۴: معماری شبکه عصبی توصیف چندوجهی

۳-۱۴-۲ شباهت زوج‌جمله - توصیف فشرده جمله

برای این بخش از کار، یک شبکه عصبی کدگذار-کدگشا طراحی کردیم که مبتنی بر کانولوشن یک‌بُعدی با فیلترهای مستقل از هم است. در این معماری ابتدا توسط عملیات کانولوشن تعداد ویژگی‌های ورودی را فشرده می‌کنیم و در بخش کدگشا سعی می‌شود تا سیگنال ورودی بازسازی شود. برای این عملیات معماری تصویر ۳-۵ استفاده شده است. در معماری ما با توجه در هر ردیف بردار ویژگی یک واژه قرار داده شده است.



تصویر ۳-۵: ساختار سیستم پیشنهادی رمزگذار - رمزگشا

همان‌طور که در تصویر ۳-۵ آمده است مراحل به صورت زیر است:

۱. Conv1D: ۱۶ پنجره پیچشی که برخلاف کانولوشن معمولی، وزن‌های هرکدام متفاوت است استفاده شده است. طول پنجره هرکدام سه در یک است.

۲. Batch Normalization & Dropout: در این مرحله ۲۵ درصد نورون‌ها از محاسبات حذف می‌شوند و مقادیر خروجی از پنجره‌ها به صورت دسته‌ای نرمال‌سازی می‌شوند، این عملیات به‌منظور جلوگیری از بیش‌برازش استفاده می‌شود.

۳. Conv1D: این‌بار از پنجره‌های بزرگ‌تر پنج در یک به‌منظور استخراج ویژگی‌های کلی‌تر استفاده شده است.

۴. Average Pooling: این عملیات به‌منظور استخراج ویژگی‌های مهم از تصویرهای تولید شده لایه

قبل و کاهش ابعاد است.

۵. Dense: معماری این شبکه از نوع نورون‌های تمام متصل است و اندازه آن به صورت دلخواه مانند شبکه Siamese استاندارد به ۵۰ تنظیم شده است.

۶. Dense: در این لایه به تعداد کل ویژگی‌های بردار ورودی نورون‌های تمام متصل تنظیم می‌شود و مقدار خروجی این نورون‌ها با مقدار ورودی برای محاسبه خطا استفاده می‌شود.

۳-۱۴-۳ شباهت زوج‌جمله - توصیف موجک

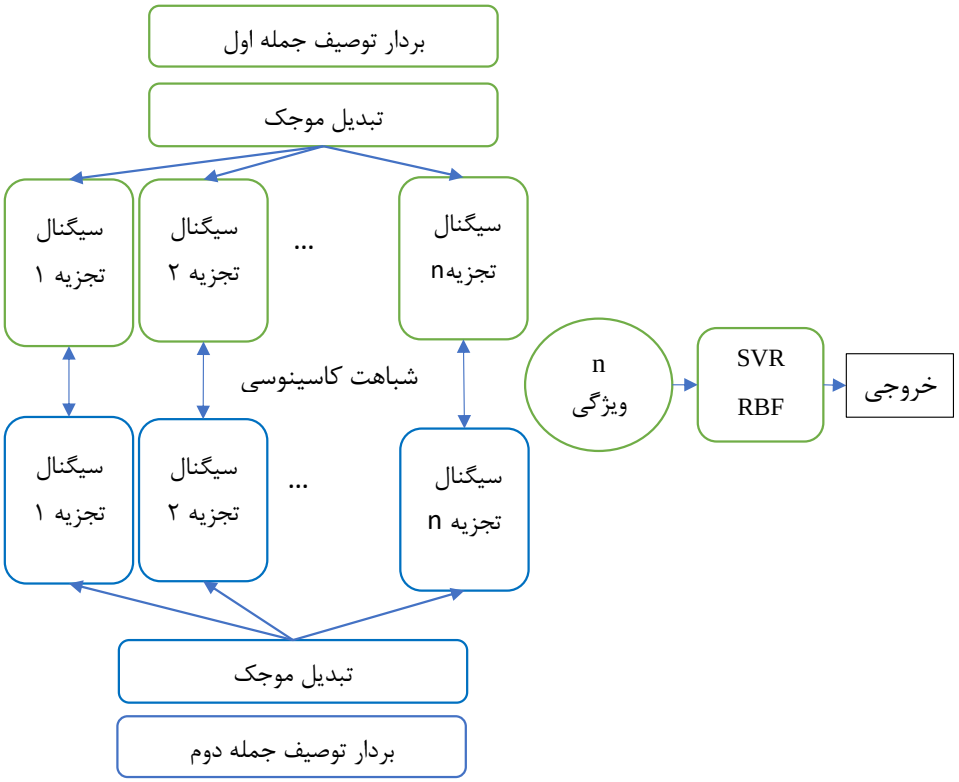
سیگنال توصیف جمله توسط تبدیل موجک به دسته‌ای از موجک‌ها تجزیه می‌شود .موجک‌ها در مقیاس زمانی به صورت محلی عمل می‌کنند و شامل دو ویژگی مقیاس و مکان هستند. مقیاس مقدار کشیدگی اندازه موجک را تعیین می‌کند و مکان تعیین می‌کند که تابع موجک در مکان خاص شامل چه مقداری است. در تبدیل موجک با عملیات پیچشی، میزان همپوشانی تابع موجک با سیگنال ورودی تعیین می‌شود، به عبارتی چقدر از تابع خاصی از موجک در سیگنال ورودی وجود دارد.

در این روش پیشنهادی، ابتدا توصیف‌گر جمله توسط رویکرد RoBERTa به دست می‌آید، اما می‌توان از هر توصیف‌گری به جای آن استفاده کرد. در مرحله بعدی همانطور که در شکل ۳-۶ نیز آمده است توصیف جمله توسط تبدیل موجک به n سیگنال تجزیه می‌شود. در مرحله بعد، شباهت کاسینوسی بردارهای متناظر تجزیه شده بین دو جمله محاسبه می‌شود و شباهت‌های محاسبه شده به عنوان ویژگی به یک رگرسیون ارسال می‌شود تا با استفاده از داده‌های آموزش مقدار شباهت را تخمین بزند. پارامترهای موجک مورد استفاده پنج کانال از نوع دبوشی نوع دو است (db2).

در جدول ۳-۴ روش‌های پیشنهادی و مشخصه آن‌ها به صورت خلاصه نمایش داده شده است.

۳-۱۵ جمع‌بندی

در روش‌های پیشنهادی معیارهای مبتنی بر شبکه واژگان موفقیت بالایی در ارزیابی شباهت داشتند. محاسبه وابستگی بین معیارهای لی، کوتاه‌ترین مسیر وزنی (صفحه ۳۷) و معیار ووپالمر باعث بهبود نتایج



تصویر ۳-۶: معماری شبکه عصبی روش توصیف موجک

هرکدام از این معیارها به صورت مستقل شده است. در روش آماری اطلاعات هم‌رخدادی مشخص شد که اهمیت بیشتر به واژگان نزدیک‌تر زوج‌واژه مورد بررسی، باعث بهبود اندکی نسبت به حالتی می‌شود که از بسامد رخداد نسبی برای محاسبه شباهت استفاده شود. استفاده از ماتریس شباهت زوج‌واژه موفق بوده است اما هنوز باید روش اصلاح شود به صورتی که امکان منفی بودن جمله در آن بررسی شود و همچنین بتواند شباهت بیشتر از دو واژه را نیز به صورت جدا محاسبه کند. همچنین عدم استفاده از روش‌های یادگیری ماشین باعث مشکل در تعمیم روش‌های پیشنهادی به پایگاه داده‌های دیگر خواهد شد. استفاده از یک روش آماری و روش Word2Set باعث پوشش بهتر نقش‌واژه‌های مختلف جمله می‌شود و همچنین واژگانی که در شبکه واژگان نباشند پوشش می‌دهد. همچنین معماری استاندارد Siamese در این فصل معرفی شد و بحث شد که توصیف‌گرهای پیشین شامل Word2Vec و Glove را می‌توان با توصیف‌گرهای دیگر جایگزین کرد. همچنین یک سیستم مبتنی بر تبدیل موجک روی توصیف جمله پیشنهاد شده است که می‌تواند بسط جزئی‌تری از بردار کل جمله فراهم نماید.

جدول ۳-۴: خلاصه‌ای از سیستم‌های پیشنهادی و مشخصه آن‌ها

روش پیشنهادی	نوع	مشخصه
شباهت تلفیقی زوج‌واژگان بخش ۳-۱۳-۱	شباهت زوج‌واژگان	Word2Set و کمیت قدرت کل
Weighted بخش ۳-۱۳-۲	شباهت زوج‌جمله	تطابق رشته‌ها، هم‌رخدادی درجه دو و معیار لی
Relative بخش ۳-۱۳-۳	شباهت زوج‌جمله	هم‌رخدادی درجه دو و شبکه واژگان
GLP بخش ۳-۱۳-۴	شباهت زوج‌جمله	تطابق‌رشته‌ها شبکه واژگان پیچیدگی جمله
WordnetPPDB بخش ۳-۱۳-۵	شباهت زوج‌جمله	شبکه واژگان WordNet شبکه واژگان قابل جایگزین PPDB
توصیف چندوجهی VSD بخش ۳-۱۴-۱	شباهت زوج‌جمله	ViCo، Dict2Vec، SynGCN
توصیف فشرده جمله بخش ۳-۱۴-۲	شباهت زوج‌جمله	معماری کدگذار- کدگشا
توصیف موجک بخش ۳-۱۴-۳	شباهت زوج‌جمله	تبدیل موجک و شباهت بردارهای تجزیه

فصل ۴: ارزیابی روش‌های پیشنهادی

۴-۱ شباهت زوج‌واژگان

۴-۱-۱ داده‌ها

پایگاه داده RG [۱۱۰] شامل ۶۵ زوج‌واژه نام است که موضوع آن‌ها از زوج‌های مترادف تا نامرتبط متغیر می‌باشد. دو گروه، در کل ۵۱ نفر از دانشگاهیان، به عنوان داوران تخصیص نمره‌های شباهت انتخاب شدند. عدد نهایی ارزیابی با میانگین‌گیری محاسبه شده است. محدوده نمره‌ها بین صفر و چهار می‌باشد. نمونه داده‌های آن را در جدول ۴-۱ مشاهده می‌کنید.

جدول ۴-۱: پایگاه داده زوج‌واژگان RG

واژه اول	cord	brother	gem
واژه دوم	smile	lad	jewel
شباهت	۰/۰۲	۲/۴۱	۳/۹۴

داده‌های MC شامل ۳۰ جفت واژه قرض گرفته شده از RG است. ۱۰ عدد از آن‌ها بسیار شبیه، ۱۰ جفت متوسط شبیه و ۱۰ جفت نامرتبط هستند. این پژوهش ۲۵ سال بعد انجام شده است و توسط ۳۸ داور دانشجو، نمره‌ها دوباره مورد ارزیابی قرار گرفته‌اند. نمونه نمره‌های جدید به شرح جدول ۴-۲ است.

جدول ۴-۲: پایگاه داده زوج‌واژگان MC

واژه اول	cord	brother	gem
واژه دوم	smile	lad	jewel
شباهت	۰/۱۳	۱/۶۶	۳/۸۴

در WS353 [۳۷] جفت‌واژگان از ۲۷ دامنه دانش شامل علوم کامپیوتری، تجاری و سرگرمی می‌باشند. ارزیابی توسط ۱۶ داور انجام شده است (۱۰-۰). نمونه‌هایی از داده‌ها را در جدول ۴-۳ مشاهده می‌کنید. داده‌های SCWS [۵۳] شامل زوج‌واژگان نام، صفت و فعل است. واژگان به‌وسیله یک توزیع، به نحوی

جدول ۴-۳: پایگاه داده زوج‌واژگان WS353 ارزیابی شباهت

واژه اول	tiger	food	month
واژه دوم	cat	rooster	hotel
شباهت	۷/۳۵	۴/۴۲	۱/۸۱

انتخاب شدند که هم شامل واژگان متداول و هم نادر باشند. به هر زوج‌واژه، توسط ۱۰ داور انسانی نمره تخصیص داده شده است. نمونه آن در جدول ۴-۴ دیده می‌شود.

جدول ۴-۴: داده‌های زوج‌واژگان SCWS

واژه اول	Brazil	Harvard	Wednesday
واژه دوم	triple	Cambridge	weekday
شباهت	۰	۷	۱/۸۱

داده‌های RW شامل ۲۰۳۴ واژه کم‌رخداد می‌باشد که توسط ۱۰ داور، به محدوده (۱-۰) ارزیابی شده است [۸۰]. نمونه آن را در جدول ۴-۵ مشاهده می‌کنید.

جدول ۴-۵: داده‌های زوج‌واژگان RW

واژه اول	undated	imperfection	elector
واژه دوم	undatable	state	voter
شباهت	۵/۸۳	۰/۲۹	۷/۸۳

در پژوهشی WS353 به دو بخش تقسیم شده است. یک پایگاه داده آن شامل ۲۰۳ زوج‌واژه که نمره‌های شباهت تخصیص داده شده است WSS و یکی دیگر شامل ۲۵۲ زوج‌واژه که نمره‌های تخصیص داده شده، بیانگر میزان ربط دو جمله است (WSR). به یاد داشته باشید که دو واژه ممکن است مرتبط ولی متضاد باشند [۳].

زوج‌واژگان MEN که حاوی ۳۰۰۰ زوج‌واژه است میزان ربط واژگان را ارزیابی می‌کند. مقادیر ربط توسط ۵۰ کارگر شرکت آمازون تخصیص داده شده است [۲۱] (به جدول ۴-۶ مراجعه شود).

داده YP-130 شامل زوج‌های نقش‌واژه فعل است که از آزمون‌هایی مانند «تافل» و «انگلیسی زبان دوم» استخراج شده است. شش داور نمره تخصیص داده‌اند که مقیاس آن از نامرتبط تا «کاملاً وابسته و مرتبط» متغیر است [۱۳۹] (جدول ۴-۷).

جدول ۴-۶: داده‌های زوج‌واژگان MEN

واژه اول	bakery	sun	beach
واژه دوم	zebra	sunlight	sand
شباهت	۰	۱	۰/۹۶

جدول ۴-۷: داده‌های زوج‌واژگان YP-130

واژه اول	brag	twist	imitate
واژه دوم	boast	curl	highlight
شباهت	۴	۳/۱۶۷	۰/۱۶۷

داده‌های MTURK-287، به‌طور خودکار از واژگان هم‌رخداد انتخاب شده است و روش ارزیابی، مشابه با روش WS353 می‌باشد. نمونه داده‌های آن در جدول ۴-۸ دیده می‌شود.

جدول ۴-۸: داده‌های زوج‌واژگان MTurk-287

واژه اول	episcopal	summer	money
واژه دوم	russia	winter	quota
شباهت	۲/۷۵	۴/۳۷۵	۲/۵

داده‌های MTURK-771 شامل موارد متنوعی از انواع ارتباطات در شبکه واژگان است [۴۴] (جدول ۴-۹).

جدول ۴-۹: داده‌های زوج‌واژگان MTurk-771

واژه اول	agency	amusement	agent
واژه دوم	police	athletics	spy
شباهت	۳/۱۹	۲/۶	۴

زوج‌نام‌های Rel-122 شامل محدوده کاملی از واژگان «کاملاً نامرتبط» تا «قویاً مرتبط» می‌باشد. واژه‌ها از شبکه‌های معنایی، مانند شبکه واژگان استخراج می‌شود. میزان ربط، توسط ۲۰ داور انسانی تعیین شده است [۱۱۸] (جدول ۴-۱۰).

جدول ۴-۱۰: داده‌های زوج‌واژگان Rel-122

واژه اول	tuition	ethanol	hotel
واژه دوم	fee	benzene	bibliography
شباهت	۳/۸۴	۲/۹۹	۰/۳۶

داده‌های Verb-143 شامل ۱۴۳ زوج‌فعل است که ۱۲۲ عدد از آن‌ها ریشه مشترک دارند. موضوع آن شامل سندهای حقوقی - قضایی می‌باشد. واژه‌ها به نحوی انتخاب شده‌اند که حداقل، مقداری شباهت داشته باشند. ده داور انسانی بین (۴-۰) نمره ارزیابی تخصیص داده‌اند [۱۰] (جدول ۴-۱۱).

جدول ۴-۱۱: داده‌های زوج‌واژگان Verb-143

واژه اول	brag	hinder	shake
واژه دوم	boast	assist	swell
شباهت	۴	۲	۰/۱۶۷

داده‌های شباهت SimLex-999 از بانک اطلاعات انجمن‌های آزاد دانشگاه فلوریدای جنوبی [۹۲] استخراج شده است. از شرکت‌کنندگان خواسته شده است که بعد دیدن یک واژه ، اولین چیزی که به ذهنشان می‌رسد را بنویسند. بنابراین تعداد دفعاتی که یک واژه در ذهن مخاطبان تداعی شده، به عنوان معیار شباهت استفاده شده است [۴۹] (جدول ۴-۱۲).

جدول ۴-۱۲: داده‌های زوج‌واژگان SimLex-999

واژه اول	old	quick	cloud
واژه دوم	new	rapid	weather
شباهت	۰	۹/۸۴	۳/۱۵

۴-۱-۲ آمار و ارقام داده‌های شباهت زوج‌واژگان

در جدول ۴-۱۳ اطلاعات اولیه‌ای در ارتباط با داده‌های مورد استفاده برای ارزیابی شباهت زوج‌واژگان نمایش داده شده است. پیکره متنی مورد استفاده برای محاسبات، پیکره بزرگ گوگل است که در بخش ۴-۳ به آن اشاره شد. در این جدول، میانگین معکوس سند اشاره به میانگین بسامد معکوس اسناد[1] دارد؛ به این معنی که میانگین بسامد رخداد معکوس (تعداد صفحه‌های حاوی واژه) بین تمام واژگان پایگاه داده متناظر محاسبه شده است. همچنین انحراف معکوس سند[2] به معنی انحراف معیار مقادیر بسامد معکوس است؛ و بنابراین، بین تمام واژگان پایگاه داده متناظر محاسبه شده است. میانگین و انحراف رخداد[3] نیز به معنی میانگین و انحراف معیار بین مقادیر «بسامد کل» رخداد است. نکته برجسته در جدول، این است که با توجه به مقادیر IDF (بسامد معکوس سند)، واژگان از تمامی احتمالات رخداد ممکن پشتیبانی می‌کند. داده‌های Rel-122، RG، MC و RW تقریباً از واژه‌های کم‌رخدادتر نسبت به

[1] Mean of inverse document frequencies

[2] Standard deviation of inverse document frequencies

[3] Mean and standard deviation of term frequencies

سایر پایگاه‌های داده تشکیل شده است. یک رابطه خطی بین تعداد واژگان یکتا و تعداد کل آن‌ها در یک پایگاه داده وجود دارد. به نظر می‌آید می‌توان گفت که داده‌های Rel-122 نسبت به MC کم‌رخدادتر هستند و تنوع بسامد رخداد آن‌ها کمتر از باقی پایگاه داده‌هاست.

آمار داده‌های WS353 و SCWS به هم شبیه است، هرچند که تعداد زوج‌واژگان موجود در هرکدام بسیار متفاوت است. به‌طور خلاصه، هر پایگاه داده امضای خود را دارد چرا که از دامنه‌های دانش متنوع و در زمان متفاوتی جمع‌آوری شده است. بعضی از واژه‌های شبکه واژگان تشکیل شده‌اند و بنابراین سازگاری بالایی برای استفاده از آن‌ها، توسط روش‌های مبتنی بر شبکه واژگان وجود دارد. این در حالی است که بعضی دیگر، از پیکره‌های متنی جمع‌آوری شده است که برای روش‌های مبتنی بر شبکه واژگان چالش‌برانگیز هستند.

جدول ۴-۱۴ نتایج آماری اولیه را روی ویژگی‌های داده‌های آموزشی نمایش می‌دهد. همان‌طور که در فصل قبل بخش ۳-۱۳-۱ اشاره شد، داده‌های آموزشی، زوج‌واژگان W1500 هستند. مقادیر آماری این جدول روی تمامی زوج‌واژگان محاسبه شده است و مقادیر کمینه، بیشینه، میانگین و انحراف معیار آن‌ها را نمایش می‌دهد. از این جدول می‌توان برداشت کرد که مقیاس مقادیر در هر ویژگی، متفاوت می‌باشد و به همین دلیل، نرمال‌سازی آن‌ها در محدوده صفر و یک الزامی است. مقدار بزرگ انحراف معیار، در مقایسه با میانگین نشان می‌دهد که دامنه مناسبی از واژگان برای تولید پایگاه داده استفاده شده است. در ویژگی کمیت قدرت کل بسامد کل، میانگین بسیار از مقدار بیشینه کمتر است و با توجه به انحراف معیار آن، به این معنی است که بیشتر مقادیر بین میانگین و $\frac{max}{2}$ متغیر هستند.

۴-۲ شباهت زوج‌جمله‌گان

STSS65 ۴-۲-۱

چندین پایگاه داده برای ارزیابی و مقایسه روش‌های پیشنهادی و پیشین استفاده شده است. داده‌های STSS65 حاوی ۳۰ زوج‌جمله آزمون و ۳۵ زوج آموزش می‌باشد؛ موضوع این جمله‌ها مطالب عمومی

جدول ۴-۱۳: پایگاه داده‌ها برای آزمودن شباهت و ربط زوج‌واژگان

پایگاه داده	نوع	تعداد زوج‌واژگان	واژگان یکتا	میانگین رخداد	انحراف رخداد	میانگین معکوس سند	انحراف معکوس سند
RG	شباهت	۶۵	۴۸	۸،۵۵۹،۵۸۶	۱۰،۸۵۳،۰۱۹	۱،۳۵۳،۶۴۵	۹۸۹،۱۴۰
MC	شباهت	۳۰	۳۹	۱۱،۰۷۶،۰۴۰	۱۲،۶۰۴،۰۰۹	۱،۵۷۹،۲۰۴	۹۸۱،۷۶۶
WS353	هردو	۳۵۳	۴۳۷	۲۷،۵۱۰،۰۷۳	۳۸،۸۵۲،۰۱۰	۱،۹۹۷،۳۹۵	۱،۳۰۷،۴۶۷
SCWS	شباهت	۱،۹۹۷	۱۷۱۳	۲۹،۵۳۷،۸۳۱	۴۹،۸۱۳،۶۹۳	۲،۱۱۸،۵۹۵	۱،۲۷۸،۲۵۷
RW	شباهت	۲،۰۳۴	۲،۹۵۱	۱۰،۹۶۷،۲۴۰	۲۹،۹۲۵،۹۹۱	۲۹۴،۲۲۰	۵۳۵،۱۸۱
WSS	شباهت	۲۰۳	۲۷۷	۲۵،۹۹۲،۸۷۶	۳۷،۵۶۷،۰۷۲	۱،۹۵۸،۷۲۸	۱،۲۵۷،۷۵۲
WSR	ربط	۲۵۲	۳۴۶	۲۹،۵۹۸،۸۰۰	۳۸،۹۱۱،۹۳۱	۲،۱۱۱،۶۳۲	۱،۳۰۲،۲۱۰
MEN	ربط	۳،۰۰۰	۷۵۱	۱۸،۴۵۵،۵۳۵	۳۷،۷۳۵،۵۳۰	۱،۶۰۲،۶۷۳	۱،۲۲۵،۱۹۱
YP-130	شباهت	۱۳۰	۱۴۷	۱۴،۶۰۹،۳۶۵	۲۸،۸۷۸،۲۸۹	۱،۵۳۰،۴۷۶	۱،۰۶۵،۴۱۰
MTURK-287	ربط	۲۸۷	۴۹۹	۱۵،۴۵۳،۸۳۲	۲۶،۲۵۴،۱۷۸	۱،۲۹۵،۲۱۲	۱،۰۳۸،۱۴۸
MTURK-771	ربط	۷۷۱	۱،۱۱۳	۲۲،۰۷۴،۱۰۵	۳۲،۶۷۲،۸۹۷	۱،۸۸۵،۳۷۰	۱،۱۶۷،۸۶۵
Rel-122	ربط	۱۲۲	۲۴۹	۶،۸۰۶،۰۶۷	۱۰،۱۹۴،۰۹۰	۹۴۰،۰۸۵	۸۱۴،۷۲۷
Verb-143	شباهت	۱۴۳	۱۴۷	۱۳،۰۸۷،۲۲۹	۲۶،۸۰۷،۲۵۹	۱،۵۳۰،۴۷۶	۱،۰۶۵،۴۱۰
SimLex-999	شباهت	۹۹۹	۱،۰۲۸	۲۷،۶۲۷،۹۴۰	۴۵،۷۱۵،۵۲۵	۲،۲۵۴،۱۹۱	۱،۳۱۱،۱۳۷

جدول ۴-۱۴: ویژگی‌های آماری پایگاه داده W1500

ویژگی	کمینه	بیشینه	میانگین	انحراف
کمیت قدرت کل بسامد کل بخش ۳-۳	۰	۰/۰۳۸	۰/۰۱	۰/۰۳
کمیت قدرت کل بسامد صفحه	۰	۰/۱۴۱	۰/۰۲۶	۰/۰۲۲
مدل سه‌نویسه‌ای گوگل بخش ۳-۴	۰	۰/۳۴۷	۰/۱۴۷	۰/۰۸۷
Word2Set بخش ۳-۵	۰	۱	۰/۱۷۴	۰/۲۱

است. طول هر جمله می‌تواند کوتاه (حدود چهار واژه) یا بلند (حدود ۳۳ واژه) باشد. برای هر زوج‌جمله، داوران انسانی نمره شباهت در بازه صفر و یک اختصاص داده‌اند. صفر به معنی «نبود شباهت زوج‌جمله» و یک بیانگر این است که دو جمله «معنی یکسانی» دارند. برای مثال، از داده‌های آزمون این مجموعه داده، می‌توان به جدول ۴-۱۵ اشاره کرد [۹۶].

روش طراحی شده، باید بتواند نمراتی را به‌طور خودکار تخصیص دهد که همبستگی بالا با نمره‌های داوران انسانی داشته باشند.

STSS131 ۲-۲-۴

مجموعه داده STSS131 شامل ۳۰ زوج‌جمله آزمون است؛ در نگاه اول نوع جمله‌ها متفاوت به نظر می‌رسد، در بین زوج‌جمله‌ها نسبت به پایگاه‌داده قبلی، واژگان مشترک بیشتری هست اما برخلاف وجود آن‌ها، دو جمله مفاهیم متفاوتی را بیان می‌کنند. موضوع جمله در این داده‌ها، مطالب عمومی می‌باشد. برای مثال، از داده‌های آزمون می‌توان به جدول ۴-۱۶ اشاره کرد. بنابراین، حتی برای روش‌هایی که

جدول ۴-۱۵: نمونه داده‌های پایگاه داده $STSS665$

شناسه	جمله اول	جمله دوم	نمره
۱	Cord is strong, thick string	A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly	۰/۰۱
	طناب رشته‌ای قوی و ضخیم است	لبخند حالتی در چهره شما است در زمانی که شما خوشحال هستید یا سرگرم شدید و یا سعی می‌کنید دوستانه رفتار کنید	
۲	A cock is an adult male chicken	A rooster is an adult male chicken	۰/۸۶۳
	یک خروس یک حیوان نر بالغ است	یک خروس یک حیوان نر بالغ است	

نتایج خوبی روی داده‌های STSS65 کسب کردند دست‌یابی به نتایج مشابه در مورد STSS131 مشکل‌تر است [۹۶].

جدول ۴-۱۶: نمونه داده‌های پایگاه داده $STSS131$

شناسه	جمله اول	جمله دوم	نمره
۱	There was a heap of rubble left by the builders outside my house this morning.	Sometimes in a large crowd accidents may happen, which can cause deadly injuries	۰/۰۲
	ساختمان‌سازان حجم زیادی خرده‌سنگ‌های ساختمانی را بیرون خانه من جا گذاشتند	گاهی در بین جمعیت زیاد حوادثی اتفاق می‌افتد که ممکن است باعث آسیب‌های کشنده‌ای شود	
۲	I am so hungry I could eat a whole horse plus dessert	I could have eaten another meal, I’m still starving	۰/۷۶۵
	آن‌قدر گرسنه هستم که می‌توانم یک اسب کامل به همراه یک دسر بخورم	من می‌توانستم یک وعده دیگر بخورم، هنوز گرسنه هستم	

SICK ۳-۲-۴

زوج‌جمله‌های SICK[۴] خود متشکل از دو مجموعه داده هستند [۸۲] که در ادامه، نوع آن‌ها توضیح داده شده است. شامل ۱۰۰۰ عدد زوج‌جمله می‌باشد. در جدول ۴-۱۷ نمونه‌هایی از این مجموعه داده ربط دو جمله آورده شده است.

۱. جمله‌های توصیف‌کننده «دو تصویر شبیه به هم» که از مجموعه ۸۰۰۰ جمله استخراج شده‌اند

[4]Sentences involving compositional knowledge (SICK) dataset

[۵۱]. توصیف‌ها در مورد تصاویر متنوع و موقعیت‌های مختلف هستند ولی در مورد اشخاص نیستند.

۲. متون توصیف ویدیوها که در آن از کاربران خواسته شده است با دیدن یک ویدیو هرکدام توصیف خود را بنویسند [۲۶].

جدول ۴-۱۷: نمونه داده‌های پایگاه داده $SICK$

شناسه	جمله اول	جمله دوم	نمره
۱	A group of kids is playing in a yard and an old man is standing in the background	There is no boy playing outdoors and there is no man smiling	۳/۳
	گروهی از کودکان در حیاط بازی می‌کنند و مردی لبخند می‌زند.	هیچ پسری بیرون بازی نمی‌کند و هیچ مردی لبخند نمی‌زند	
۲	A fearful little boy is on a climbing wall	The man is sitting next to a birdcage	۱/۱
	یک پسربچه ترسیده از روی دیوار بالا می‌رود	مردی کنار قفس پرنده ایستاده است	

Stsbenchmark ۴-۲-۴

این پایگاه داده ارزیابی شباهت، خود از چندین زیر مجموعه تشکیل شده است[۵] که در ادامه لیست آن‌ها آورده شده است. در جدول ۴-۱۸ نمونه‌هایی از این داده‌ها را مشاهده می‌کنید [۲۵].

- MSRP: پیکره متنی که شامل زوج‌جمله‌هایی است که یک خبر را با واژگان متفاوت توصیف می‌کنند [۶].
- headlines: مهم‌ترین اخبار از رسانه‌های اروپایی[۷]
- deft-news: مقاله‌های خبری
- MSR-Video: زیرنویس گفتار اخبار ویدیویی.

[5]STSbenchmark (2017). STS Wiki [online]. website: http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark [ed 15 July 2021]

[6]Microsoft research paraphrase corpus

[7]European media monitor

- ImageDescription: شامل توصیف اشخاص، خودروها، حیوانات و اشیاء داخل خانه و اداره است [۱۸].

- track5.en-en: استخراج شده از پیکره استنتاج زبان طبیعی استنفورد [۳۳] [8].

- answers-answers و anwer-forums: از وب‌سایت StackExchange [9]

جدول ۴-۱۸: نمونه داده‌های پایگاه داده $Stsbenchmark$

شناسه	منبع	جمله اول	جمله دوم	نمره
۱	MSRVid	A plane is taking off.	An air plane is taking off.	۵
		هواپیما شروع به پرواز کرد	هواپیما شروع به پرواز کرد	
۲	Images	A passenger train in the snow.	A passenger train waiting in a station.	۲/۲
		قطار مسافری در برف.	قطار مسافری که در ایستگاهی منتظر است.	
۳	deft-news	Zahedan is a city in southeastern	Bam is a city in southeastern	۱/۸
		زاهدان شهری در جنوب شرق است	بم شهری در جنوب شرق است	
۴	headlines	Indonesian president to visit UK	Indonesian president to visit Australia	۱/۴
		رئیس‌جمهور اندونزی به بریتانیا سفر می‌کند	رئیس‌جمهور اندونزی به استرالیا سفر می‌کند	

در ادامه، ابتدا روش‌های پیشنهادی را با یکدیگر مقایسه می‌کنیم؛ سپس بهترین آن‌ها در قیاس با روش‌های پیشین، ارزیابی می‌شود. پس از آن، روش‌های پیشنهادی، از دیدگاه چهار معیار آماری مقایسه شده‌اند.

۴-۳ معیارهای ارزیابی

در تمامی معیارهای ارزیابی، نماد X به معنای بردار مقادیر پیش‌بینی شده شباهت، Y نماد مقادیر بردار هدف و نمادهای $\bar{X}$ و $\bar{Y}$ میانگین عناصر بردار هستند.

[8] Stanford natural language inference (SNLI)

[9] StackExchange (2005). stack exchange [online]. website: https://stackexchange.com [accessed 20 June 2020]

۴-۳-۱ همبستگی پیرسون و اِسپیرمَن

این دو مهم‌ترین معیارهای مقایسه دو روش مبتنی بر رگرسیون هستند، هرچند که همبستگی پیرسون [۱۰] نسبت به اسپیرمن[۱۱] معیار مهم‌تری برای ارزیابی شباهت در نظر گرفته می‌شود [۴۷] اما فاصله اسپیرمن نسبت به نویز حساسیت کمتری دارد [۴۷]. نحوه محاسبه معیار پیرسون در فرمول ۴-۱ آورده شده است:

$$pr = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (۱-۴)$$

برای محاسبه معیار اسپیرمن فرمول ۴-۲ استفاده شده است [۱۳۵].

$$sp = 1 - \frac{cov(rg_X, rg_Y)}{\sigma_{rg_x}\sigma_{rg_x}} \quad (۲-۴)$$

در این فرمول، rg_x بردار رتبه[۱۲] مقادیر هدف است و rg_y بردار رتبه پیش‌بینی می‌باشد. نماد cov در این فرمول به معنای کوواریانس، σ_{rg_x} به معنای انحراف معیار مقادیر بردار X و σ_{rg_y} نماد انحراف معیار مقادیر بردار Y است.

در توضیح بیشتر می‌توان گفت که ابتدا مقادیر بردار X به صورت صعودی مرتب می‌شوند و به عناصر بردار، شماره شاخصی از یک تا n که تعداد عناصر بردار است اختصاص داده می‌شود. در ادامه، مقادیر پیش‌بینی شده، رو به روی مقدار متناظر هدف قرار می‌گیرند و شماره شاخص آن‌ها با توجه به تغییر مکانشان تنظیم می‌شود. بدین صورت دو بردار rg_x و rg_y شکل می‌گیرد. در مرحله بعد، با نماد d^2 مقدار

[10] Pearson

[11] Spearman

[12] Rank

تفاوت این دو بردار رتبه، به توان دو، محاسبه می‌شود. بدین صورت، می‌توان فرمول ۴-۲ را به ۴-۳ تغییر داد.

$$sp = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (۳-۴)$$

۴-۳-۲ میانگین انحراف

این معیار[۱۳] از فرمول زیر محاسبه می‌شود:

$$md = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \quad (۴-۴)$$

هرقدر اختلاف مقادیر تخصیص داده شده در میانگین انحراف با نمره‌های داوران انسانی کمتر باشد، این معیار کوچک‌تر می‌شود و در نتیجه روش پیشنهادی، بهتر است.

۴-۳-۳ خطای استاندارد باقی‌مانده (RSE)

باوجود استخراج ویژگی‌ها، حتی اگر بهترین پارامترهای روش را بدانیم، بازهم نمی‌توانیم رگرسیون را کاملاً دقیق پیش‌بینی کنیم؛ دلیل این موضوع در نظر نگرفتن بعضی ویژگی‌های پنهان یا وجود نویز در اندازه‌گیری ویژگی‌ها می‌باشد. می‌توان با استفاده از معیار خطای استاندارد باقی‌مانده [۱۴] مقدار این ناهماهنگی را محاسبه کرد. این معیار از فرمول ۴-۵ به دست می‌آید [۶۳]:

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (۵-۴)$$

[13] Mean deviation

[14] Residual standard error

۴-۳-۴ خطای آماری R^2

این معیار[15]، معیاری دیگر برای اندازه‌گیری فاصله بین رگرسیون پیش‌بینی‌شده و مقادیر هدف است. تفاوت آن با خطای استاندارد باقی‌مانده این است که مستقل از مقیاس نمره‌های داوران انسانی می‌باشد. این معیار، محاسبه می‌کند که چه میزان تغییرپذیری مقادیر واقعی با رگرسیون پیش‌بینی‌شده توضیح داده‌شده است. زمانی که مقدار یک شود، به معنی همبستگی کامل مقادیر دو بردار است. این معیار مطابق فرمول زیر محاسبه می‌شود [۶۳]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - X_i)}{\sum_{i=1}^{n}(Y_i - \bar{Y_i})} \quad (۶-۴)$$

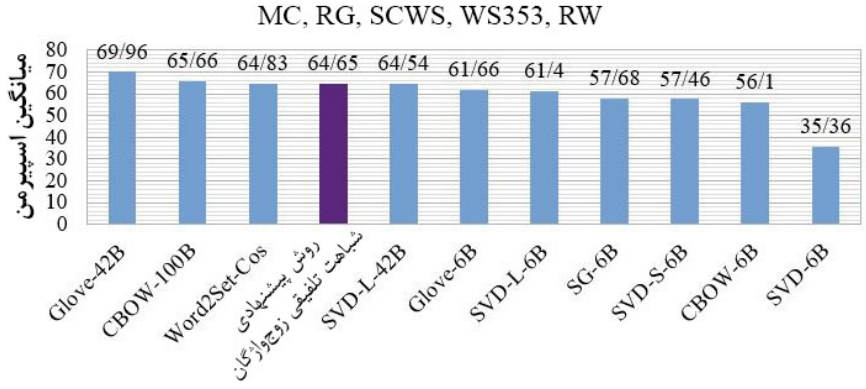
۴-۴ آزمون روی داده‌های زوج‌واژگان

تصویر ۴-۱ نتایج بین روش شباهت تلفیقی زوج‌واژگان را با روش‌های پنینگتون [۱۰۱] و خمینز [۶۶]، از دیدگاه میانگین همبستگی اسپیرمن روی داده‌های MC, RG, WS353, SCWS و RW نمایش می‌دهد. در لیست زیر روش‌ها و منابع آن‌ها آمده است:

SG-6B[۱۰۱]	SVD-S-6B[۳۰ ،۱۰۱]	CBOW-6B[۸۵ ،۱۰۱]	SVD-6B[۳۰ ،۱۰۱]
Word2Set[۶۶]	SVD-L-42B[۳۰ ،۱۰۱]	Glove-6B[۱۰۱]	SVD-L-6B[۳۰ ،۱۰۱]
		Glove-42B[۱۰۱]	CBOW-100B[۸۵ ،۱۰۱]

نتایج نشان می‌دهد مدل Glove که توصیفگر آن مبتنی بر هم‌رخدادی واژگان است، به‌طور میانگین بهترین نتایج را کسب کرده است. روش CBOW (مبتنی بر Word2Vec) در این بین، رتبه دوم را کسب کرده است، در حالی که CBOWروی ۱۰۰ میلیارد، و Glove روی ۴۲ میلیارد واژه آموزش دیده است. بنابراین، بیشتر بودن تعداد واژگان (برای آموزش) در روشی متفاوت لزوماً به معنای بهتر بودن آن نیست. همچنین، مقایسه‌ها با روش Word2set بیانگر شبیه بودن نتایج به روش پیشنهادی شباهت

[15] R^2 statistics



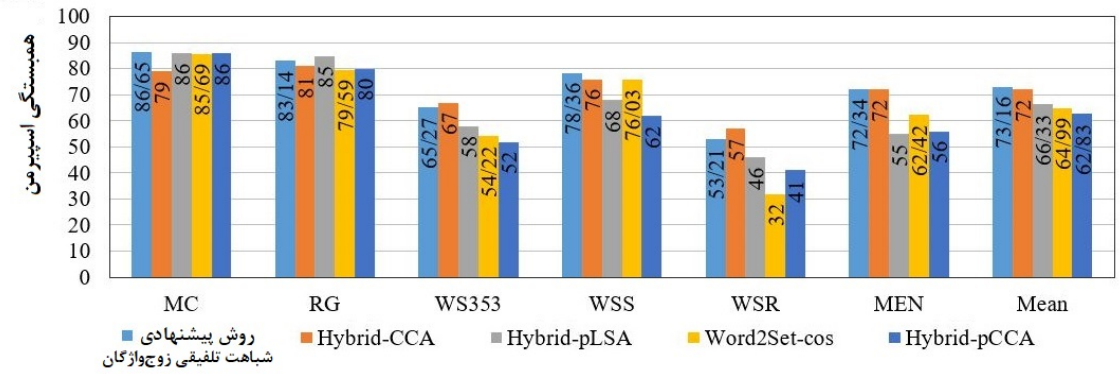
تصویر ۴-۱: مقایسه روش شباهت تلفیقی زوج‌واژگان با روش پنینگتون [۱۰۱] و خمینز [۶۶] از دیدگاه میانگین همبستگی اسپیرمن روی داده‌های MC, RG, WS353, SCWS و RW.

تلفیقی زوج‌واژگان بود. Glove و CBOW نسبت به اندازه داده‌های آموزشی حساس بوده‌اند چراکه نسبت همبستگی آن‌ها با تعداد شش میلیارد واژه کاهش یافته است. می‌توان گفت که برتری نتایج آن‌ها، به دلیل حل مشکل پراکندگی داده‌ها است. این مدل‌ها آموزش دیده‌اند تا واژه‌ها را در یک توصیف فشرده با توجه به واژگان همسایه توصیف کنند؛ در مقایسه، روش شباهت تلفیقی زوج‌واژگان از اطلاعات همسایگی کمتری استفاده می‌کند اما عملکرد آن قابل رقابت است.

در نتیجه مواردی که اشاره شد، مسئله‌ای که همچنین می‌تواند نتایج SVD را تشریح کند، این است که در رویکرد SVD، مدل توسط اطلاعات ماتریس هم‌رخدادی با ۱۰۰۰۰ واژه (X_{trunc}) پرتکرار محاسبه می‌شوند. SVD-L از $\sqrt{X_{trunc}}$ و SVD-S از $log(1 + X_{trunc})$ محاسبه شده است. SG در واقع روش اسکیپ‌گرام Word2Vec است که واژگان همسایه را بر مبنای واژه مرکزی تخمین می‌زند. درازای سرعت بالاتر این روش، کارایی در آن کاهش یافته است.

تصویر ۴-۲، نتایج روش شباهت تلفیقی زوج‌واژگان را با روش‌های «التراس و استیونسون [۶]» و «خیمنز [۶۶]» از دیدگاه همبستگی اسپیرمن مقایسه می‌کند. مدل Hybrid-CCA مبتنی بر مدل توزیعی واژگان هم‌معنی است که میانگین وزنی این مترادف‌ها یک توصیف بهتر برای واژه هدف تشکیل می‌دهند. تحلیل همبستگی متعارف[16] با تعیین و تغییر دو بردار توصیفگر واژه، باعث می‌شود زوج‌واژگان همبستگی بیشتری نسبت به حالت عادی داشته باشند [۴۵]. منابع روش‌های مورد مقایسه در ادامه

[16] CCA



تصویر ۴-۲: مقایسه روش پیشنهادی با روش‌های التراس و استیونسون [۶] و خیمنز [۶۶] از دیدگاه همبستگی اسپیرمن

آمده است:

Hybrid-pCCA[۶] Hybrid-pLSA[۶] Word2Set[۶۶] Hybrid-CCA[۶]

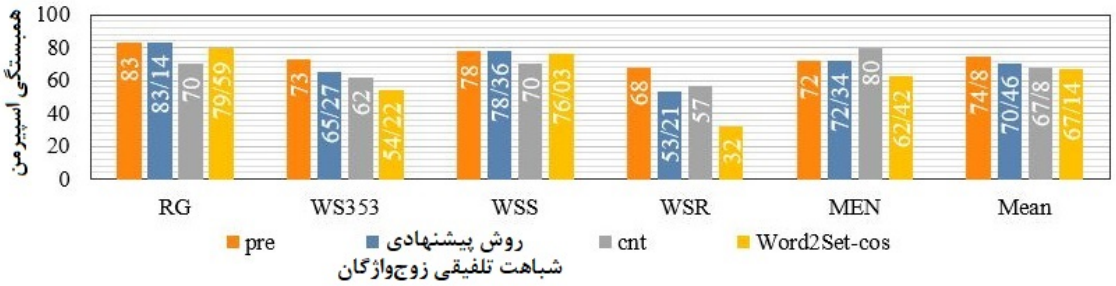
در رویکرد Hybrid-p، مدل رتبه‌بندی مترادف‌ها[17] برای محاسبه «میانگین وزنی بردار توصیفگر واژه» استفاده شده است [۴۸]. این بردار، خود مبتنی بر الگوریتم رتبه‌بندی شخصی‌سازی شده PPR[18] می‌باشد. SRM وزن واژه‌های همسایگی مرتبط را که از شبکه واژگان به دست می‌آورد بیشتر می‌کند؛ طوری که اگر نامرتبط باشند، سهم کمتری در تولید این بردار وزنی داشته باشند. بعد این عملیات، آنالیز معنای نهفته و تحلیل همبستگی متعارف برای توصیف فشرده‌تر و معنادارتر واژه استفاده می‌شوند.

از نمودار، می‌توان دریافت که روش «شباهت تلفیقی زوج‌واژگان» به‌طور میانگین (Mean)، بهتر از روش‌های دیگر مورد مقایسه می‌باشد. بهترین نتایج، روی داده‌های MC و WSS کسب شده است. این تصویر از چندین نظر درخور توجه است، زیرا ابتدا نشان می‌دهد که روش شباهت تلفیقی و Hybrid-CCA نتایج پایدارتری روی داده‌های مختلف کسب کرده‌اند. دوم، نتایج ما همیشه بهترین یا دومین بهترین بوده است و سوم، نتایج نشان می‌دهد که اضافه کردن اطلاعات آماری به شبکه واژگان Yago (مورد استفاده ما)، باعث بهبود می‌شود و البته بهترین نتایج نیز در زمینه ارزیابی شباهت به دست آمده است.

نتایج نشان می‌دهد که همه روش‌ها همبستگی بسیار خوبی روی داده‌های MC و RG به دست

[17] Synset page rank model (SRM)

[18] Personalized page rank algorithm



تصویر ۴-۳: مقایسه روش پیشنهادی با روش‌های پژوهش بارُنی [۱۱]

آورده‌اند. یافته‌ها بیانگر عملکرد خوب همه روش‌ها در ارزیابی شباهت مفاهیم عمومی می‌باشد. نکته جالب در مورد این نمودار، عملکرد بهتر روش شباهت تلفیقی نسبت به روش‌های Word2Set هم در مورد MEN و همچنین WSR است که هردو از داده‌های ارزیابی ربط دو واژه هستند.

تصویر ۴-۳ مقایسه روش شباهت تلفیقی با روش‌های بارُنی [۱۱] را نمایش می‌دهد. pre، مدل بردارهای از پیش‌آموزش دیده Word2Vec است، که در پنجره پنج واژه همسایگی آموزش دیده است تا توصیفگر ۴۰۰ بُعدی بسازد. cnt، مدل مبتنی بر اطلاعات متقابل نظیر به نظیر [19] است [۹۴] و تفاوت آن با روش مدل مبتنی بر اطلاعات متقابل نظیر به نظیر استاندارد آن است که اهمیت بیشتری به واژگان هم‌رخداد در سند ورودی می‌دهد [۵]. منابع روش‌های مورد مقایسه در ادامه آمده است:

Word2Set[۶۶] cnt[۱۱] pre[۱۱]

روش‌های آماری، در ارزیابی ربط واژگان، عملکرد بهتری نسبت به سایر روش‌ها مانند Word2Set خیمنز [۶۶] داشته‌اند. روش شباهت تلفیقی به خوبی روش خیمنز روی داده‌های MEN عمل کرده است؛ اگرچه، فاصله زیادی بین روش cnt و سایر روش‌ها در مورد RG و WSS وجود دارد. همچنین روش شباهت تلفیقی نتایج بهتری نسبت به Word2Set کسب کرده است، این در حالی است که Word2Set نتایج نوسان‌داری روی داده‌های ربط زوج‌واژه کسب کرده است.

به‌طور میانگین، بهترین عملکرد را مدل pre داشته است. Word2Set و cnt عملکردی مشابه و متوسط داشته‌اند، در حالی که روش تلفیقی ما (آماری و شبکه واژگان) نزدیک‌ترین عملکرد را نسبت به روش‌های شبکه عصبی داشته است.

[19] Point-wise mutual information

جدول ۴-۱۹: مقایسه روش پیشنهادی با روش‌های پیشین از دیدگاه درصد همبستگی اسپیرمن

نام دادگان	روش پیشنهادی (r)	بهترین روش (r)	W2S-SVR (r)	منبع بهترین روش
MC	۸۵/۶۵	۹۱	۸۵/۰۷	Patwardhan و Pedersen (۲۰۰۶) [۹۸]
YP-130	**۸۳/۴۴**	۷۴/۷	۷۳/۸۵	روش شباهت تلفیقی زوج‌واژگان، قبلاً [۴۲]
RG	۸۳/۱۴	۹۰	۷۲/۷۳	Patwardhan و Pedersen (۲۰۰۶) [۹۸]
Verb-143[۲۰]	**(۸۶/۸۱ ،۸۲/۸۶)**	(نامعلوم، ۶۴/۲)	(۳۲/۸۰، نامعلوم)	روش شباهت تلفیقی زوج‌واژگان، قبلاً [۱۰]
WSS	۷۸/۳۶	۸۰	۷۲/۳۳	Baroni و همکاران (۲۰۱۴) [۱۱]
MEN	۷۲/۳۴	۸۰	۶۱/۴۰	Baroni و همکاران (۲۰۱۴) [۱۱]
WS353	۶۵/۲۷	۸۱	۵۶/۸۸	Halawi و همکاران (۲۰۱۲) [۴۴]
MTURK-287	۶۲/۰۷	۷۳/۷	۴۵/۲۴	Halawi و همکاران (۲۰۱۲) [۴۴]
SCWS	۵۷/۳۶	۶۱/۰۴	۵۷/۰۳	Li و همکاران (۲۰۱۴) [۷۴]
MTURK-771	۵۷/۳۳	۷۲/۷	۵۴/۴۷	Halawi و همکاران (۲۰۱۲) [۴۴]
WSR	۵۳/۲۱	۷۲	۳۶/۴۷	Agirre و همکاران (۲۰۰۹) [۳]
Rel-122	۵۳/۱۸	۵۳/۴	۴۷/۱۱	Szumlanski و همکاران (۲۰۱۳) [۱۱۸]
SimLex999	۵۰/۳۹	۵۲	۵۳/۲۷	Jimenez و همکاران (۲۰۱۹) [۶۶]
RW	۳۰/۸۳	۴۷/۸	۴۱/۷۵	Pennington و همکاران (۲۰۱۴) [۱۰۲]

جدول ۴-۱۹، روش پیشنهادی را با بهترین نتایج روش‌های مختلف و همچنین بهترین روش خیمنز [۶۶] که Word2Set-SVR است، مقایسه کردیم.

از جدول ۴-۱۹ می‌توان دریافت که روش شباهت تلفیقی زوج‌واژگان در هردو YP-130 و Verb-143 بیشترین همبستگی را با نمره‌های انسانی داشته است. برداشت ما از دلیل آن، این است که دو پایگاه داده یاد شده بیشتر حاوی زوج‌فعل‌ها هستند ولی شبکه واژگان بیشتر شامل نقش‌واژه‌های نام می‌شود. ما این مشکل را با افزودن آمار مبتنی بر متن حل کردیم که موجب بهبود قابل‌توجه در نتایج شد. مدل بِیْکِر و همکاران نیز نتایج قابل توجهی در ارزیابی شباهت زوج‌فعل‌ها داشته است [۱۰]. مدل بدون ناظر آن‌ها، نمونه فعل‌های مشابه را به خوشه‌های معنایی مشترک نگاشت می‌کند. در روش حاج‌طیب و همکاران، ریشه واژگان با استفاده از الگوریتم ساده پورتر[۲۱] استخراج می‌شود [۴۲]. سپس برای یک زوج‌واژه، طبقه متناظر در گراف واژگان لینک شده بهم، در ویکی‌پدیا یافت می‌شود و ربط بین دو واژه توسط نرخ رده‌بندی‌های مشترک بین دو واژه مشخص می‌شود.

نتایج جدول ۴-۱۹ را می‌توان با داده‌های تصاویر پیشین مقایسه کرد و به این نتیجه رسید که روش پیشنهادی نلفیقی در زمینه ارزیابی شباهت روی داده‌های MC, RG, WSS و YP-130 بهتر کار کرده است. به‌طور عمومی این روش، نتایج نویدبخشی در مورد شباهت زوج‌فعل‌ها کسب کرده است اما در

[21] Porter

ارتباط با واژگان نادر مانند RW عملکرد بهتری مورد نیاز است. ما فرض کردیم می‌توان با افزودن واژگان بیشتر از طریق مدل سه‌نویسه‌ای گوگل، نتایج را بهبود بخشید اما در عمل، با افزودن آن خطای حداقل مربعات افزایش یافت و در پی آن مقادیر همبستگی کاهش داشت. بنابراین ما بیشتر این روش را توسعه ندادیم.

۴-۴-۱ ارزیابی اهمیت ویژگی‌های روش پیشنهادی - شباهت تلفیقی زوج‌واژگان

در جدول ۴-۲۰ ویژگی‌های مورد استفاده برای محاسبه شباهت/ربط زوج‌واژگان به طور مستقل مورد ارزیابی قرار گرفته‌اند. نتایج نوشته شده در این جدول، توسط روش آزمون تی‌زوج نمونه‌ها[۲۲] روی داده‌های آموزشی W1500 محاسبه شده است.

جدول ۴-۲۰: اهمیت و اولویت هر ویژگی با توجه به مقدار متناظر p-value

ویژگی	نتیجه آزمون تی	نتیجه با شرط ($p < 0.05$)	اولویت ویژگی
کمیت قدرت کل-بسامد کل	$1/75497946e-52$	رد	3
کمیت قدرت کل-بسامد صفحه	$1/13450986e-80$	رد	2
مدل سه‌نویسه‌ای گوگل	0	رد	4
Word2Set-Cos	$6/09421867e-92$	رد	1

ابتدا باید اشاره کرد که ما همبستگی اسپیرمن ۰/۸۰ را با اعتبارسنجی متقابل 10-fold با استفاده از فرآیند رگرسیون گاوسین[۲۳] به دست آوردیم. با هدف کاهش خطای جذر مربعات خطا با نمره‌های انسانی، ویژگی «کمیت قدرت کل مبتنی بر بسامد کل» فقط یک‌بار انتخاب شد، در حالی که سایر ویژگی‌ها همیشه انتخاب شدند. تصویر ۴-۲۰ نتایج ارزیابی اهمیت هرویژگی، نسبت به مقدار هدف را نمایش می‌دهد. دقت بفرمایید که فرضیه صفر[۲۴] این است که رابطه‌ای قوی بین یک ویژگی و مقدار هدف وجود ندارد.

از نتایج این جدول می‌توان نتیجه‌گیری کرد که ما این فرضیه را برای تمامی ویژگی‌ها رد می‌کنیم زیرا همه آن‌ها مقدار $p - value < 0.05$ کسب کردند. این بدان معنی است که بین هر ویژگی و مقدار

[22] Paired sample t-test

[23] Gaussian processes regression (GPR)

[24] Null hypothesis

هدف رابطه‌ای قوی وجود دارد.

۴-۴-۲ مقدمه در ارتباط با ارزیابی سیستم‌های پیشنهادی

در ادامه، تعدادی از روش‌های پیشنهادی برای شباهت زوج‌جمله را روی داده‌های مختلف براورد می‌کنیم. در پژوهش‌های مختلف، مهمترین معیار ارزیابی در محاسبه شباهت زوج‌جمله‌ها همبستگی پیرسون در نظر گرفته می‌شود. البته در گذشته در پژوهش‌هایی [۵۹] معیارهای دیگری مانند همبستگی اسپیرمن و میانگین انحراف نیز در نظر گرفته می‌شد. پژوهش ما به دو بخش تقسیم می‌شود:

در بخش اول، برای داده‌های کوچک آموزش و آزمون، روش‌های «تلفیقی آماری و شبکه واژگان» مورد استفاده قرار می‌گیرند، بدین منظور ابتدا روش‌های Relative, WordnetPPDB, Weighted و GLP با شیوه‌های پیشین که روی داده‌های کوچک ارزیابی می‌شدند مقایسه شده است. در بخش دوم، شبکه‌های عصبی و داده‌های بزرگ مورد ارزیابی قرار گرفتند و در مورد نتایج مقایسه، بحث و بررسی صورت گرفته است.

۴-۵ آزمون روی داده‌های STSS65

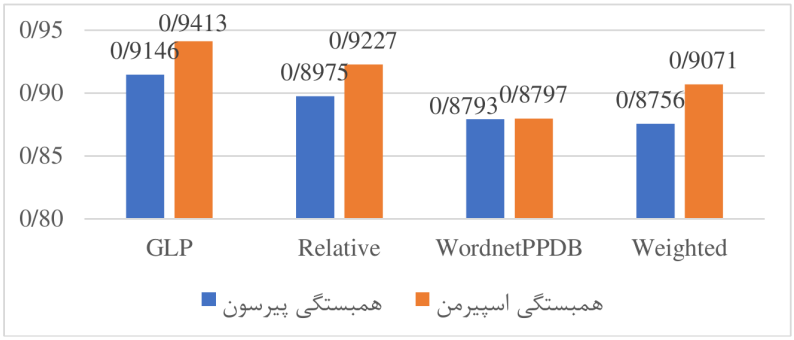
۴-۵-۱ مقایسه از دیدگاه همبستگی

مقایسه روش‌های پیشنهادی Relative, Wordnet-PPDB, Weighted و GLP با یکدیگر از دیدگاه همبستگی پیرسون در تصویر ۴-۴ آورده شده است. نتایج مقایسه روش‌های پیشین با روش پیشنهادی GLP در تصویر ۴-۵ آورده شده است. منابع رویه‌های پیشین مورد مقایسه در جدول ۴-۲۱ آمده است.

جدول ۴-۲۱: لیست منابع روش‌های پیشین در مقایسه با روش پیشنهادی GLP

LSA[۳۰]	Liu[۷۷]	STS[۵۹]	Omiotis[۱۲۴]	CTS[۵۴]
	Feng[۳۵]	SymSS[۹۵]	LSS[۲۹]	STASIS[۷۶]

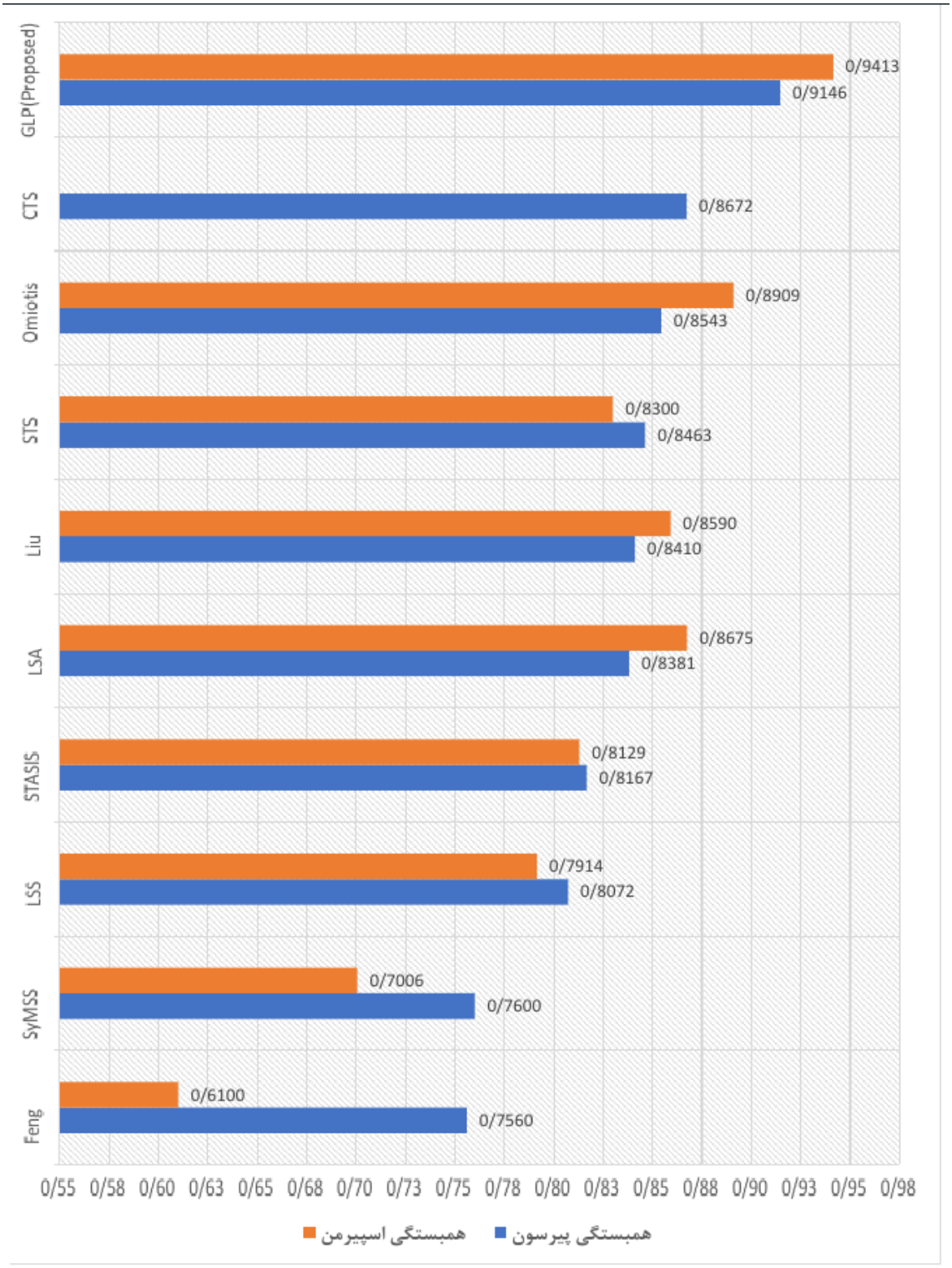
همان‌طور که در تصویر آورده شده است، روش «ویژگی‌های عمومی و محلی (GLP)» بهترین نتایج را به‌دست آورده است. این نتایج نشان‌دهنده اهمیت استخراج ویژگی‌های محلی و تلفیق آن با سایر



تصویر ۴-۴: مقایسه روش‌های پیشنهادی Relative, Wordnet-PPDB, Weighted و GLP با یکدیگر از دیدگاه همبستگی

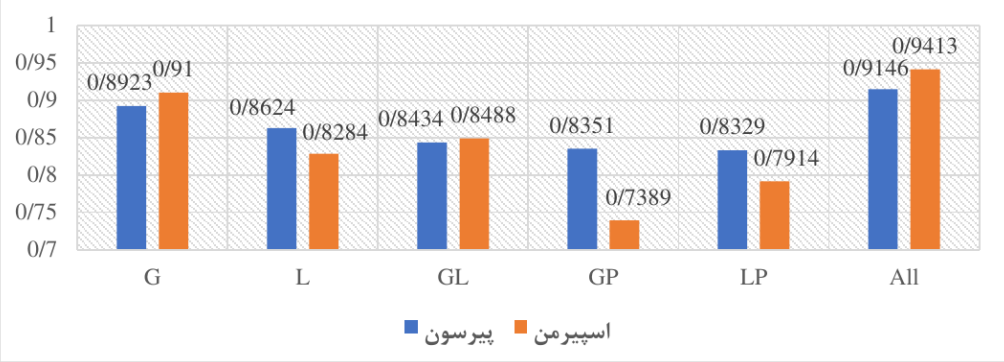
ویژگی‌ها است. همچنین استفاده از پیچیدگی جمله، باعث بهبود چند صدم درصدی روش ویژگی‌های عمومی-محلی شد. اطلاعات هم‌رخدادی در روش‌های پیشنهادی «واژگان همسایگی وزن‌دهی شده (Weighted)» و «شباهت مفهوم‌ها (Relative)» مؤثر بوده‌اند. نتایج نشان‌دهنده اهمیت بالای معرفی پارامتر وابستگی بین متغیرها برای مدل کردن غیرخطی رگرسیون است (در همه روش‌های پیشنهادی به‌غیراز Weighted).

همان‌طور که در تصویر ۴-۵ دیده می‌شود، روش پیشنهادی GLP که مبتنی بر استخراج ویژگی‌های عمومی و محلی از ماتریس شباهت زوج‌واژگان است، نسبت به سایر روش‌ها بهبود قابل‌توجهی داشته است. در روش GLP همانند رویکرد CTS محتوای اطلاعات مفهوم‌ها در محاسبه شباهت استفاده شده است. همچنین معیار «وابستگی فاصله‌ی وزنی مسیر (Weigthed Path) -لی» تأثیر قابل‌توجهی در عملکرد سیستم داشته است. روش Omiotis و STS نیز نتایج قابل‌توجهی داشته‌اند. مهم‌ترین ویژگی در ارتباط با روش STS، تکیه کامل این روش روی اطلاعات آماری محاسبه شباهت است؛ به عبارتی دیگر این روش، از اطلاعات معیارهای مبتنی بر شبکه واژگان استفاده نمی‌کند و بنابراین محاسبه شباهت نسبت به زبان، مستقل است. الگوریتم تطابق واژگان در روش پیشنهادی GLP از همین روش مشتق شده است. در روش GLP علاوه بر آن‌که برای اولین بار معیار فاصله وزنی مسیر استفاده شده است؛ همچنین پیاده‌سازی وابستگی دو معیار محاسبه شده روی شبکه واژگان Yago نیز باعث بهبود نتایج شده است. روش Omiotis با استفاده از وزن‌دهی واژگان (با توجه به نقش آن‌ها) در جمله توانسته است نتایج قابل‌توجهی کسب کند. روش آنالیز معنای نهفته (LSA) توانسته است با خوشه‌بندی واژگان



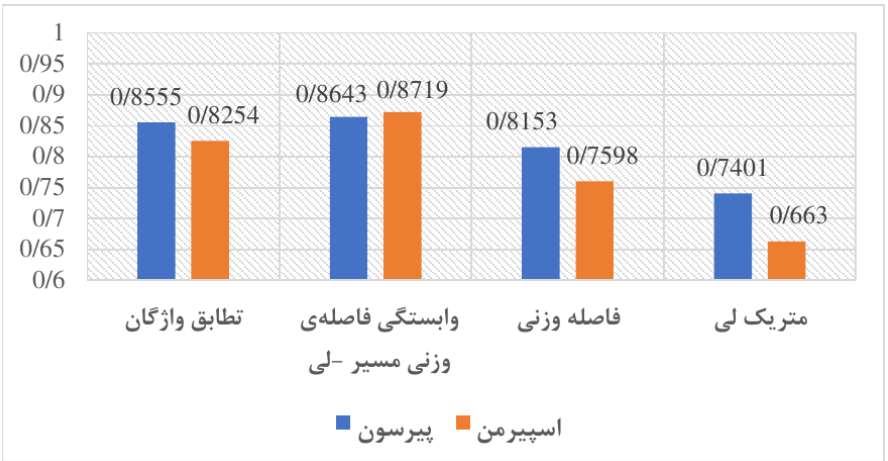
تصویر ۴-۵: مقایسه روش پیشنهادی GLP با روش‌های پیشین از دیدگاه همبستگی

شبیه به هم، بتواند نتایج خوبی در محاسبه شباهت به‌دست بیاورد. در تصویر ۴-۶، همبستگی پیرسون هریک از ویژگی‌های «عمومی (G)»، «محلی (L)»، «عمومی و محلی (GL)»، «عمومی و پیچیدگی متن (GP)»، «محلی و پیچیدگی متن (LP)» و «عمومی-محلی-پیچیدگی متن (All)» با نمره‌های داوران انسانی نمایش داده شده است.



تصویر ۴-۶: همبستگی اجزای روش پیشنهادی GLP از دیدگاه نوع ویژگی

همان‌طور که در تصویر ۴-۶ مشخص است، سه ویژگی بهترین عملکرد را در تلفیق با یکدیگر دارند. پیچیدگی متن به تنهایی نتوانسته است تأثیر چندانی در همبستگی داشته باشد و تلفیق ویژگی‌های محلی و عمومی، نسبت به تلفیق هریک با پیچیدگی متن، نتیجه بهتری دربر داشته است. همچنین بهترین نتیجه را ویژگی‌های عمومی داشته است و «تأثیر آن در تلفیق» برای بهبود نتایج قابل‌توجه می‌باشد. در تصویر ۴-۷ همبستگی پیرسون هریک از ویژگی‌های روش پیشنهادی GLP، به صورت مستقل، با نمره‌های داورانی انسانی نمایش داده شده است.

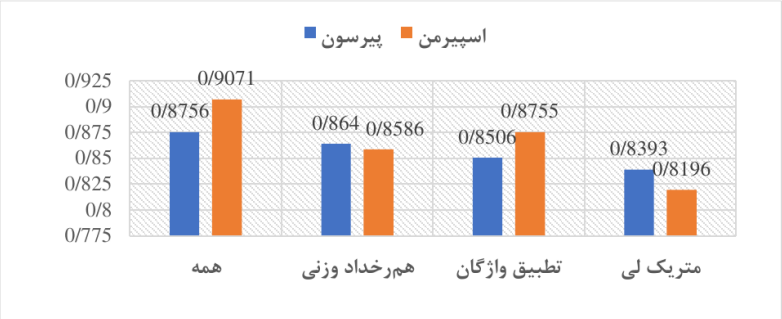


تصویر ۴-۷: همبستگی ویژگی‌های مستقل سیستم پیشنهادی GLP با داده‌های STSS65

همان‌طور که در تصویر ۴-۷ مشخص است، استفاده از یک معیار وابستگی (فاصله وزنی مسیر-لی) باعث بهبود قابل‌توجه نتایج شده است؛ همچنین روش تطابق واژگان به تنهایی نتایج خوبی به دست آورده است. بعلاوه، معیار فاصله وزنی (کوتاه‌ترین مسیر وزن‌دهی شده) نتایج بهتری نسبت به روش لی داشته است. بنابراین نتایج این فرضیه را تائید می‌کنند که استفاده از معیار وابستگی «بین دو معیار شبکه‌ی واژگان»، بعلاوه «پیچیدگی متن» و «ویژگی‌های عمومی» در محاسبه شباهت، باعث بهبود همبستگی بین نمره‌های رگرسیون پیش‌بینی و هدف می‌شوند.

۴-۵-۲ همبستگی اجزای روش واژگان همسایگی وزن‌دهی شده (Weighted)

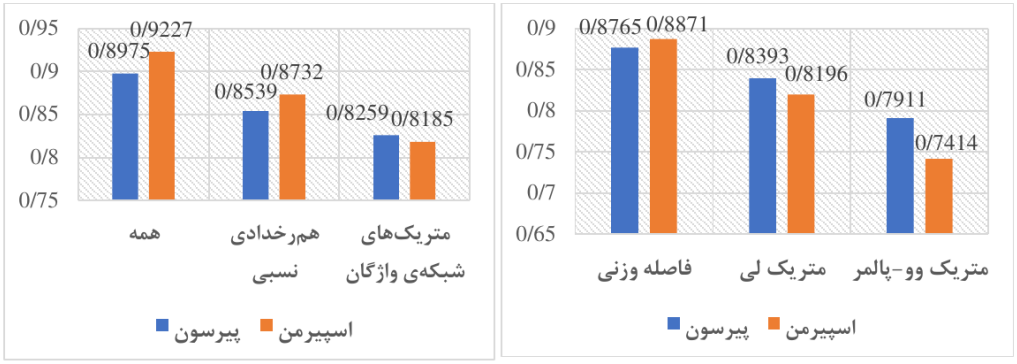
در تصویر ۴-۸ مقایسه همبستگی اجزای مختلف روش پیشنهادی Weighted به صورت مستقل آورده شده است. برای انجام این آزمون، برای هریک از ویژگی‌ها به صورت مستقل ماتریس M در نظر گرفته می‌شود. در نظر داشته باشید که در این رویکرد، فقط ویژگی‌های عمومی استخراج می‌شود. همان‌طور که در تصویر ۴-۸ آورده شده است، نتایج روش هم‌رخداد وزنی (WSOC-PMI) نسبت به بقیه کمی بهتر می‌باشد در حالی که روش لی از بقیه ضعیف‌تر بوده است. با مقایسه این تصویر با تصویر ۴-۷ می‌توان نتیجه گرفت که استفاده از ویژگی‌های عمومی و محلی در بهبود نتایج الگوریتم تطبیق واژگان، مؤثر نبوده است. اما نمی‌توان در مورد روش لی همین نتیجه را گرفت؛ چرا که روش لی با استفاده از تلفیق ویژگی‌های محلی و عمومی، توانسته است بهبود قابل‌توجهی داشته باشد؛ بنابراین میزان این تأثیر، وابسته به معیاری است که مورد استفاده قرار می‌گیرد.



تصویر ۴-۸: همبستگی ویژگی‌های مستقل سیستم پیشنهادی Weighted با داده‌های STSS65

۴-۵-۳ همبستگی اجزای روش هم‌رخدادی واژه‌ها و شباهت مفهوم‌ها (Relative)

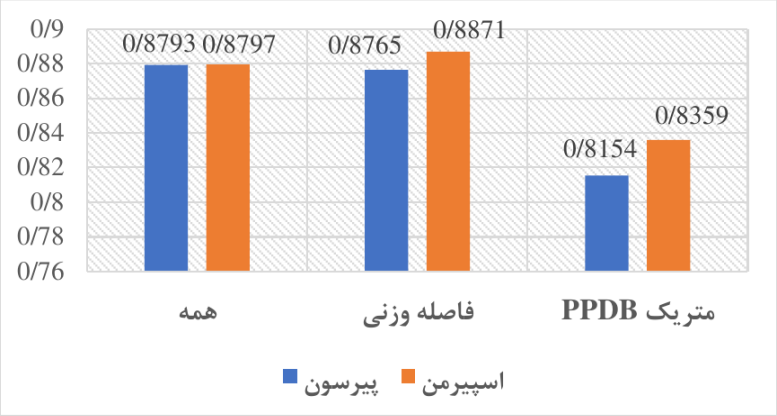
مقایسه تأثیر هریک از اجزای روش Relative در تصویر ۴-۹ آورده شده است. همان‌طور که در تصویر ۴-۹ مشخص است، معیار فاصله وزنی نتیجه بهتری نسبت به سایر معیارها دارد؛ بهتر بودن روش‌های لی و فاصله وزنی نشان‌دهنده اهمیت بالای محاسبه کوتاه‌ترین مسیر در ارزیابی شباهت می‌باشد؛ زیرا ک روش پالمر تنها روشی است که فقط عمیق مفهوم مشترک را در نظر می‌گیرد. در تصویر ۴-۹، با مقایسه نتایج روش هم‌رخدادی نسبی Relative با نوع وزنی Weighted، می‌توان نتیجه گرفت که استفاده از وزن‌های بیشتر برای واژگان نزدیک‌تر، نتایج را اندکی بهبود داده است.



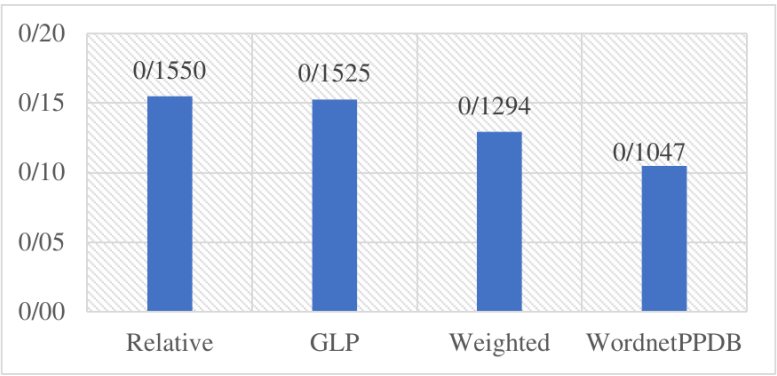
تصویر ۴-۹: همبستگی ویژگی‌های مستقل سیستم پیشنهادی Relative با داده‌های STSS65

۴-۵-۴ همبستگی ویژگی‌های روش WordNet و PPDB (WordnetPPDB)

برای این بخش، معیارهای مبتنی بر WordNet و پایگاه دانش PPDB محاسبات مستقلی دارند؛ مقادیر همبستگی هرکدام در تصویر ۴-۱۰ نمایش داده شده است. این تصویر نشان می‌دهد که معیار مبتنی بر PPDB تأثیر ناچیزی روی بهبود نتایج کل داشته است، هرچند که نتایج آن به صورت مستقل عالی ارزیابی می‌شود. استفاده از معیار PPDB بیشتر بار محاسباتی برای سیستم خواهد داشت، بنابراین به جای این روش، استفاده مستقل از فاصله وزنی پیشنهاد می‌شود.



تصویر ۴-۱۰: مقایسه اجزای روش WordNetPPDB روی داده‌های STSS65



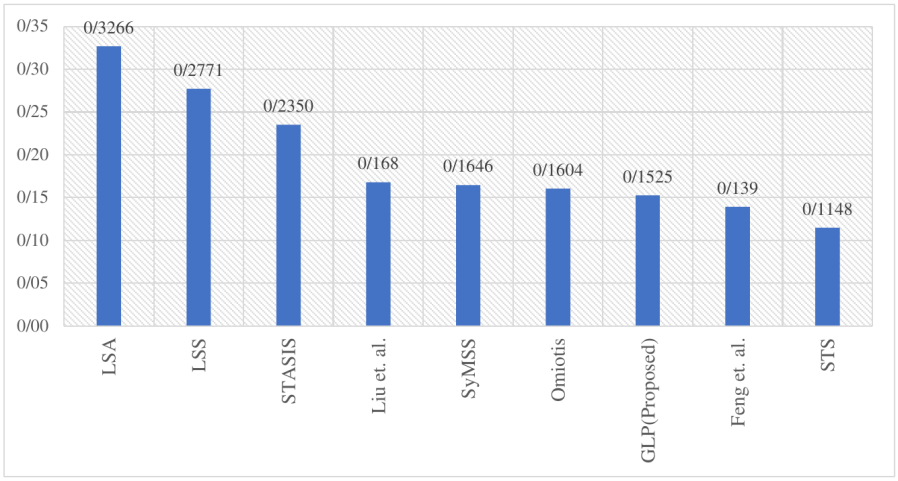
تصویر ۴-۱۱: مقایسه روش‌های پیشنهادی Relative, WordnetPPDB, Weighted و GLP از دیدگاه میانگین انحراف روی داده‌های STSS65

۴-۵-۵ مقایسه با روش‌های پیشین از دیدگاه میانگین انحراف

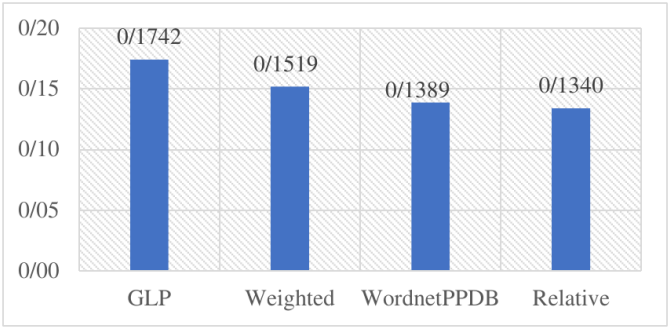
در تصویر ۴-۱۱ ابتدا روش‌های پیشنهادی Relative, WordnetPPDB, Weighted و GLP با یکدیگر نسبت به این معیار مقایسه شده‌اند.

با توجه به تصویر ۴-۱۱، روش مبتنی بر «واژگان همسایگی وزن‌دهی شده» و روش WordNetPPDB، بهترین عملکرد را در این بخش داشتند. در تصویر ۴-۱۲ چهار روش پیشنهادی یاد شده را با روش‌های دیگر مقایسه می‌کنیم.

با نگاه به تصویر ۴-۱۱ و مقایسه آن می‌توان فهمید که بین روش‌های پیشنهادی روش Weighted از روش GLP از دیدگاه میانگین انحراف (STS) نتیجه بهتری داشته است. روش GLP از روش‌های LSA, LSS و STASIS نتایج بهتری کسب کرده است و اختلاف آن، در مقایسه با روش‌های Omiotis



تصویر ۴-۱۲: مقایسه روش‌های پیشنهادی Relative, WordnetPPDB, Weighted و GLP با روش‌های پیشین از دیدگاه میانگین انحراف روی داده‌های STSS65



تصویر ۴-۱۳: مقایسه روش‌های پیشنهادی Relative, WordnetPPDB, Weighted و GLP از دیدگاه خطای استاندارد باقی‌مانده روی داده‌های STSS65

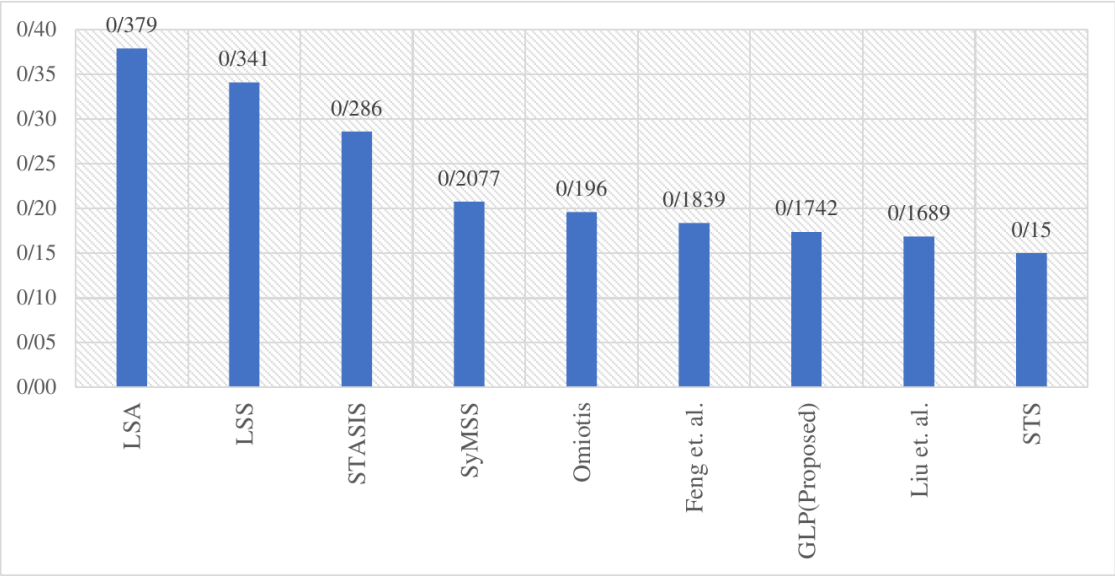
و SyMSS چشم پوشیدنی است.

۴-۵-۶ مقایسه با روش‌های پیشین از دیدگاه خطای استاندارد باقی‌مانده

در تصویر ۴-۱۳ ابتدا روش‌های پیشنهادی Relative, WordnetPPDB, Weighted و GLP با یکدیگر نسبت به این معیار مقایسه شده‌اند. هرچقدر مقدار این معیار به صفر نزدیک‌تر باشد، روش پیشنهادی بهتر است.

مقایسه چهار روش پیشنهادی از خطای استاندارد باقی‌مانده با یکدیگر نشان می‌دهد که روش Relative بهترین عملکرد را در این بخش داشته است. روش‌های مبتنی بر «هم‌رخدادی متقابل واژگان (GLP)» و (Weighted) عملکردی مشابه هم داشتند و به‌طور کلی اختلاف قابل‌توجهی بین چهار روش پیشنهادی وجود ندارد. در تصویر، مقایسه روش GLP با روش‌های پیشین نیز نمایش داده شده است.

منابع رویکردهای پیشین مورد مقایسه، در جدول ۴-۲۱ صفحه ۱۲۵ آمده است.



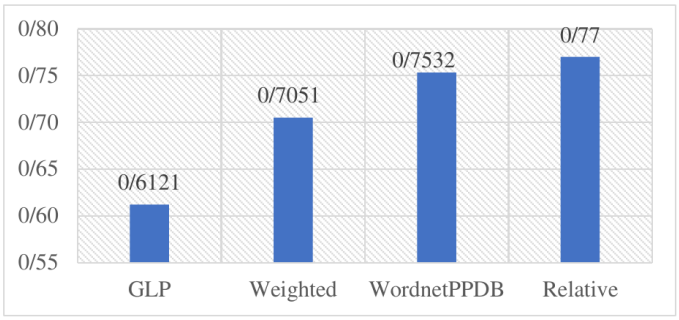
تصویر ۴-۱۴: مقایسه روش‌های پیشنهادی Relative, WordnetPPDB, Weighted و GLP با روش‌های پیشین از دیدگاه خطای استاندارد باقی‌مانده روی داده‌های STSS65

با توجه به تصویر ۴-۱۴، روش پیشنهادی GLP همراه با روش STS بهترین عملکرد را دارند و با سایر رقبا فاصله قابل‌توجهی دارند. روش‌های LSA, LSS و STASIS، همپوشانی خوبی با رگرسیون هدف ندارند و اختلاف قابل‌توجهی با سایر روش‌ها دارند. روش‌های Liu و Feng نیز نتایج خوبی به‌دست آوردند و قابل مقایسه با روش پیشنهادی و STS هستند.

۴-۵-۷ مقایسه با روش‌های پیشین از دیدگاه آماری R^2

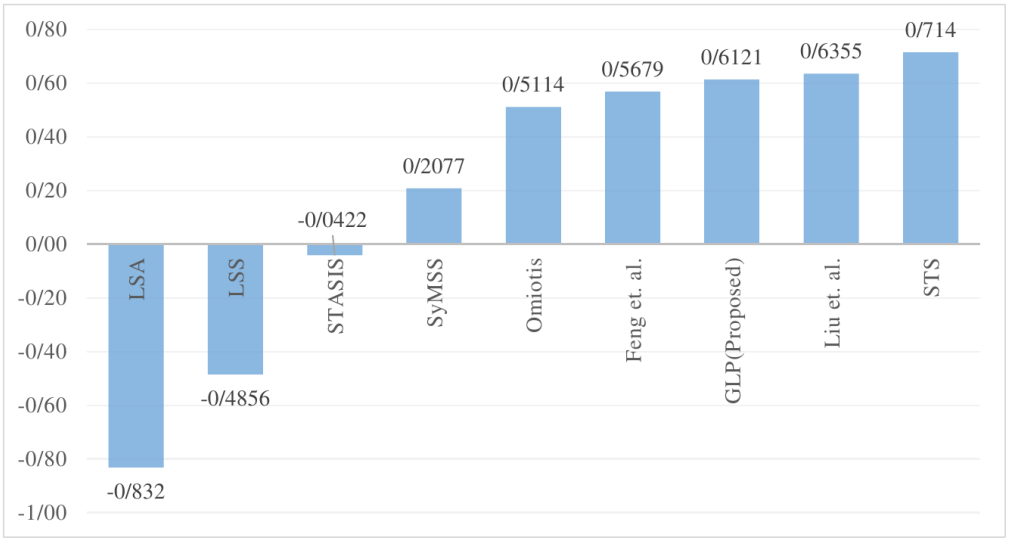
در تصویر ۴-۱۵، ابتدا روش‌های پیشنهادی Relative, WordnetPPDB, Weighted و GLP با یکدیگر نسبت به این معیار مقایسه شده‌اند. هرچقدر مقدار این معیار به یک نزدیک‌تر باشد، روش پیشنهادی رگرسیون هدف را بهتر مدل کرده است.

در این آزمون در تصویر ۴-۱۵، روش هم‌رخدادی نسبی و شباهت مفهوم‌ها (Relative) بهترین نتیجه را کسب کرده‌اند؛ و روشی که کاملاً مبتنی بر پایگاه دانش است (Weighted) نتیجه بسیار خوبی نسبت به سایر روش‌های پیشنهادی کسب کرده است. روش مبتنی بر «ویژگی‌های عمومی-محلی GLP»، فاصله قابل‌توجهی نسبت به روش‌های دیگر دارد و در این آزمون عملکرد ضعیف‌تری داشته است. در



تصویر ۴-۱۵: مقایسه روش‌های پیشنهادی Relative, WordnetPPDB, Weighted و GLP از دیدگاه R^2 روی داده‌های STSS65

تصویر ۴-۱۶ روش GLP را با روش‌های پیشین مقایسه می‌کنیم. منابع رویکردهای پیشین در جدول ۴-۲۱ صفحه ۱۲۵ آمده است.



تصویر ۴-۱۶: مقایسه روش‌های پیشنهادی با روش‌های پیشین از دیدگاه R^2

با توجه به تصویر ۴-۱۶، در مقایسه با روش‌های پیشین، روش STS بهترین نتایج را کسب کرده است، روش Liu دومین بهترین است و روش GLP در جایگاه سوم می‌باشد. البته باید این نکته را ذکر کرد که سه روش یاد شده، نسبت به روش استاندارد LSA، روش LSS, STASIS و SyMSS نتایج بسیار بهتری کسب کرده‌اند. مقدار منفی در این آزمون، بدین معناست که هرگاه مقدار رگرسیون هدف بالا رفته است مقدار پیش‌بینی‌شده در این مدل‌ها پایین رفته است و همچنین خلاف این قضیه نیز وجود دارد. روش LSA و LSS در این آزمون عملکرد بسیار ضعیفی داشته‌اند. روش Omiotis نیز نتیجه متوسطی داشته است و توانسته تنها ۵۰ درصد تغییرات رگرسیون هدف را به درستی مدل کند.

جدول ۴-۲۲: مقایسه رویکردهای پیشین با روش‌های پیشنهادی روی مجموعه داده STSS131

	ویژگی‌های عمومی-محلی	واژگان همسایگی وزن‌دهی شده	هم‌رخداد نسبی و شباهت مفهوم‌ها	WordNet PPDB	آنالیز معنای نهفته
	GLP	Weighted	Relative	WordnetPPDB	LSA
همبستگی پیرسون	0/734	0/71	0/69	0/682	0/736
همبستگی اسپیرمن	0/729	0/667	0/62	0/606	0/696
میانگین انحراف	0/147	0/144	0/155	0/153	0/144
خطای استاندارد باقی‌مانده	0/203	0/206	0/23	0/202	0/194
R^2	0/447	0/431	0/289	0/45	0/493

۴-۵-۸ آزمون روی مجموعه داده STSS131

به علت قابل دسترس نبودن کدها یا نمره‌های همبستگی روش‌های مقایسه شده در بخش‌های قبل، در این بخش نتوانسته‌ایم همه این روش‌ها را مقایسه کنیم اما سایر روش‌ها در جدول ۴-۲۲ مقایسه شده‌اند.

نتایج بیانگر آن است که روش LSA و GLP با نتایج بسیار نزدیک به هم، بهترین روش‌های حل مسئله روی این مجموعه داده هستند. همچنین روش پیشنهادی GLP از نظر همبستگی پیرسون تنها روشی است که مقدار بالای ۷۰ درصد کسب کرده است. اختلاف قابل‌توجهی از دیدگاه همبستگی پیرسون بین روش‌های مقایسه شده وجود ندارد، اما در آزمون اسپیرمن اختلاف معناداری بین روش‌ها قابل‌مشاهده است؛ روش‌های Relative و WordnetPPDB که در آن‌ها از «معیارهای وابستگی تلفیقی شبکه واژگان» استفاده شده است، اختلاف زیادی با بقیه روش‌ها دارند، اگرچه، نمی‌توان نتیجه‌گیری کرد که استفاده از این معیارها باعث اختلاف شده است. به‌طور کلی، با مقایسه نتایج روی این مجموعه داده و STSS65 به نظر می‌رسد که روش آماری STS، بهترین عملکرد را نسبت به «روش‌های مبتنی بر شبکه‌ی واژگان» در بیشتر آزمون‌ها داشته است.

نتایج همبستگی پیرسون در جدول ۴-۲۲ بیانگر آن است که استخراج ویژگی امر مهمی در محاسبه شباهت می‌باشد زیرا که روش GLP بهترین عملکرد را دارا است، هرچند که اختلاف قابل توجهی با رقبا

نداشته است. اما در ارتباط با همبستگی اسپیرمن نیز روش GLP موفق بوده است.

اما در مورد عملکرد خوب GLP، دلیل این موضوع بیشتر به دلیل موفقیت معیار وابستگی «فاصله وزنی مسیر-لی-ووپالمر» بوده است؛ چرا که اطلاعات هم‌رخدادی درجه دوم در روش Weighted و Relative نیز مورد استفاده قرار گرفته است. در ارتباط با RSE (خطای استاندارد باقی‌مانده)، می‌دانیم که مقدار «فاصله مطلق مقادیر پیش‌بینی» را از مقادیر هدف تعیین می‌کند و هرچقدر مقدار آن کمتر باشد بهتر است و آنکه تحلیل آن مشابه با MSE می‌باشد. مقادیر موجود نشان می‌دهد که روش LSA، بهتر از سایر روش‌ها توانسته است اختلاف بین مقدار پیش‌بینی شده و هدف را مدل کند؛ بنابراین این دو بردار کمترین اختلاف را با همدیگر دارند. اما در ارتباط با R^2 که درصد نسبی واریانس که توسط مدل‌ها تشریح شده را اندازه‌گیری می‌کند، می‌توان به نتایج بهتر دو روش WordnetPPDB و LSA اشاره کرد.

در مقایسه میانگین انحراف از نمره‌های داوران انسانی، تفاوت قابل‌توجهی بین روش‌ها وجود ندارد اگرچه هردو روش Weigthed و LSA در این زمینه بهترین عملکرد را داشته‌اند. همین موضوع در ارتباط با خطای استاندارد باقی‌مانده نیز صادق است. همان‌طور که در جدول مشخص است همه روش‌ها نتایج نزدیکی دارند و Relative کمی با بقیه روش‌ها، هرچند ناچیز، اختلاف دارد. در ارتباط با R^2 روش LSA بهترین نتیجه را دارد به این معنا که توانسته است در حدود ۵۰ درصد رگرسیون هدف را با مقادیر پیش‌بینی‌شده توضیح دهد. در این بخش فقط Relative اختلاف زیادی با بقیه روش‌ها دارد.

به‌طور کلی روش GLP و LSA توانسته‌اند بهترین نتایج را کسب کنند و به نظر می‌آید که برای بهبود بیشتر، روش‌های پیشنهادی بایستی غیر از شباهت‌ها، بتوانند مقدار اختلاف دو جمله را نیز مدل‌سازی کنند هرچند که این ممکن است باعث شود تا روی مجموعه داده دیگری نتایج همبستگی کاهش یابد.

۴-۵-۹ نتایج روش شبکه عصبی توصیف‌گر چندوجهی

همان‌طور که در بخش ۳-۱۴-۱ صفحه ۱۰۱ توضیح داده شد، این روش مبتنی بر تلفیق سه توصیفگر، و معماری گسترش‌یافته شبکه Siamese استاندارد[۲۵] است. توصیفگرهای مورد استفاده، خود مبتنی بر

[25] Siamese Sentence Similarity(2019). Github - MahmoudWahdan [online]. website: https://github.com/MahmoudWahdan/Siamese-Sentence-Similarity [accessed 15 July 2021]

«تصاویر نزدیک به هم، تعریف واژه‌ها در دیکشنری و درخت تجزیه وابستگی واژگان جمله» هستند.

در تصویر ۴-۱۷ می‌توان روند یکی از اجراهای آموزش Siamese استاندارد با Word2Vec را مشاهده کرد. دلیل نتایج نسبتاً ضعیف را می‌توان به دو موضوع نگاشت کرد.

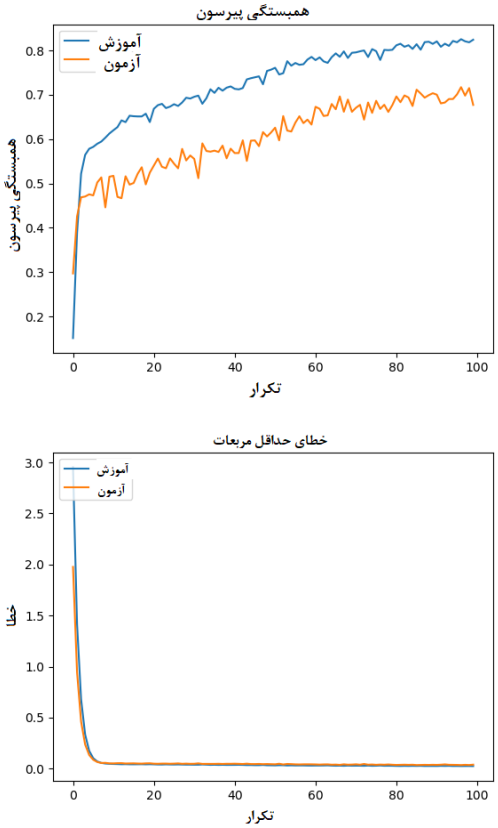
۱. پژوهش‌های پیشین که با استفاده از توصیف‌گرهای واژه آموزش دیدند، متوجه عملکرد ضعیف مدل‌های زبانی استاندارد در ارزیابی شباهت شدند. بدین معنی که توصیفگرها، برای تشکیل مدل زبانی عمومی، و نه برای مسئله خاص آموزش دیده بودند. بنابراین بردار توصیف Word2Vec استاندارد را دوباره روی داده‌های مرتبط مانند SemEval2013 آموزش داده و بهینه کردند.

۲. بعضی رویکردها از وزن‌های از پیش‌آموزش دیده برای شروع آموزش شبکه‌عصبی استفاده می‌کنند و می‌دانیم که نحوه وزن‌دهی شبکه‌های عصبی در عملکرد آن‌ها بسیار مهم است [۹۰].

اما آنچه در تصاویر ۴-۱۷ و ۴-۲۲ مشاهده می‌کنید نتایج آموزش معماری متقارن Siamese با بردار Word2Vec استاندارد است که به منظور تهیه مدل زبانی (و نه لزوما جهت ارزیابی شباهت زوج‌جمله‌های کوتاه) آموزش دیده است. نسخه استاندارد این توصیف‌گر به صورت رایگان قابل تهیه است[۲۶].

در تصویر ۴-۱۷ مشاهده می‌شود که شبکه عصبی قبل از بیش‌برازش، در حدود کمتر از ۱۰ تکرار به کمینه حداقل مربعات خطا رسیده است. منابع روش‌های مورد مقایسه در جدول ۴-۲۴ صفحه ۱۴۰ آمده است. در ارتباط با جدول ۴-۵-۹، تغییر اندازه دسته‌های آموزشی بهبودی حاصل نکرد. همچنین استفاده از Adam نسبت به AdaDelta پیشنهاد می‌شود زیرا AdaDelta بسیار کند عمل می‌کرد و در بیش از ۱۰۰ تکرار هم نتایج Adam را کسب نکرد [۹۰]. تغییر اندازه دسته‌های آموزش باعث جهش‌های منفی در آموزش می‌شد به صورتی که برای مثال بعد ۱۰ تکرار، همبستگی داده‌های آموزش جهش منفی داشت، اگرچه این مشکل با افزایش اندازه دسته‌های آموزشی بهبود پیدا کرد؛ به این دلیل که شبکه عصبی با اندازه دسته‌های کوچک‌تر، در کمینه محلی گیر می‌کند. افزودن به اندازه حافظه LSTM در تمامی توصیف‌گرها باعث شد تا خطای حداقل مربعات بیشتر شود و شبکه زودتر به بیش‌برازش برسد.

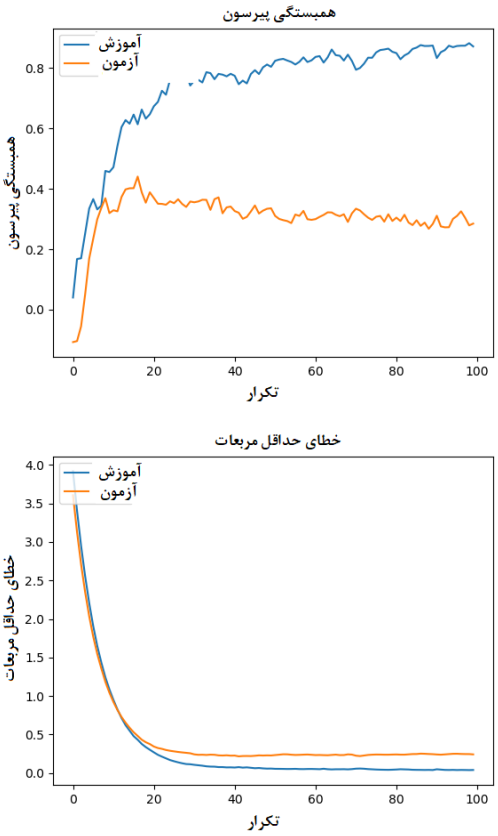
[26]Google Word2Vec(2013). Amazon Web Services [online]. website: https://s3.amazonaws.com/dl4j-distribution/GoogleNews-vectors-negative300.bin.gz [accessed 15 July 2021]



تصویر ۴-۱۷: نتایج شبکه عصبی مبتنی بر Word2Vec روی داده‌های SICK

در تصویر ۴-۲۲ نتایج همبستگی روی داده‌های SICK را بررسی کردیم. در این تصویر، کارایی توصیفگرها در شبکه Siamese مستقل و دو روش معروف Glove و Word2Vec مورد مقایسه قرار گرفته است. منابع روش‌ها در جدول ۴-۲۴ آمده است.

در جدول ۴-۲۵ نمونه‌ای از ارزیابی روش (پیشنهادی) توصیف‌گر چندوجهی روی داده‌های آزمون SICK آورده شده است. همانطور که در این جدول مشاهده می‌کنید، مقادیر پیش‌بینی شده همبستگی خوبی با بردار هدف دارند. می‌توان دید که از نظر رگرسیون، سیستم طراحی شده توان پیش‌بینی رفتار مقادیر هدف را دارا است. اما در ارتباط با روش RoBERTa-base می‌توان دید که مقیاس اعداد تولید شده بالا می‌باشد و در نتیجه میانگین خطا بالا انتظار می‌رود. دلیل این موضوع، عدم آموزش شبکه روی مجموعه داده‌های ارزیابی ربط است. زوج‌جمله‌های استفاده شده ساختار واژگان مشابهی دارند، گاهی ممکن است تعدادی واژه متفاوت باشد اما مقدار ربط دو جمله پایین باشد چرا که موضوع به کار رفته متفاوت است. تشخیص این موضوع حتی در توصیف‌گرهای وابسته به محتوا مانند RoBERTa-base



تصویر ۴-۱۸: نتایج شبکه عصبی مبتنی بر Word2Vec روی داده‌های Stsbenchmark

آنهم بدون پیش‌آموزش روی این سبک مجموعه داده، مشکل است.

همانطور که در ادامه و در روش‌های دیگری مانند BERT نیز بررسی شده است، مسئله پیش‌آموزش و شیوه شروع وزن‌دهی شبکه از نکات بسیار مهم در آموزش آن است. همچنین به دلیل پیچیدگی محاسباتی بالا، نیاز به سخت‌افزارهای قوی و حجم بالای مدل، معمولاً دو نوع مدل base و large آموزش داده می‌شود که دومی مدلی پیچیده‌تر و اما دقیق‌تر است. در ارتباط با پارامترهای مورد استفاده برای آموزش شبکه VSD می‌توان به جدول ۴-۵-۹ اشاره کرد. در توجیه اهمیت بهینه بودن پارامترهای شبکه و مقادیر بردار توصیف‌گر می‌توان به تصاویر ۴-۱۹ و ۴-۲۰ اشاره کرد، چرا که توصیف‌گرهای مورد استفاده، روی داده‌های آموزش پایگاه داده متناظر بهینه شده‌اند. در ادامه، این مسئله بیشتر مورد بحث قرار گرفته است.

تارنمای paperswithcode[۲۷] عملکرد بهترین روش‌های ارائه شده روی مجموعه داده SICK را

[27]Papers with code (2021). Semantic Textual Similarity on SICK [online]. website: https://paperswithcode.com/sota/semantic-textual-similarity-on-sick [accessed 15 July 2021]

جدول ۴-۲۳: پارامترهای بهینه برای شبکه عصبی Siamese

پارامتر	مقدار تنظیم شده
اندازه دسته‌ها	۱۲۸
بهینه‌ساز	Adam
تعداد تکرار	۱۳
تابع خطا	حداقل خطای مربعات
تعداد ابعاد توصیف‌گر	۳۰۰
نوع حافظه مورد استفاده	LSTM
اندازه حافظه	۵۰
معیار فاصله در لایه آخر	فاصله L1-norm
تابع نورون خروجی	سیگموید

جدول ۴-۲۴: لیست منابع روش‌های پیشین در مقایسه با روش پیشنهادی VSD

Dict2Vec[۱۲۲]	SynGCN[۱۲۸]	ViCo[۴۱]	Word2Vec[۸۵]	GloVe[۱۰۱]
XLM[۷۱]	SMART-RoBERTa[۶۵]	StructBERT[۱۲۹]	MNet-Sim[۶۴]	SimCSE[۳۹]
			SRoBERTa-NLI[۱۰۹]	RoBERTa[۱۰۹]

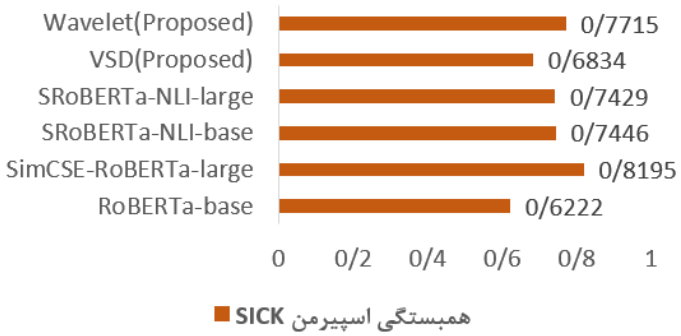
به‌روزرسانی می‌کند.

همانطور که در تصویر ۴-۱۹ مشخص شده است، مدل RoBerta-base مانند روش Word2Vec (در دو تصویر ۴-۲۲ و ۴-۲۳) عملکرد نسبتا ضعیفی داشته است و بعد بهینه‌سازی توصیف‌گر برای حل مسئله استنتاج زبان طبیعی[۲۸]، روش SRoBERTa-NLI تشکیل شده است که نسبت به حالت

[28]Natural language inference

جدول ۴-۲۵: نمونه آزمون روش پیشنهادی VSD و نتایج آن

شناسه	جمله اول	جمله دوم	مقدار هدف	مقدار پیش‌بینی	مقادیر RoBERTa-base
۱	No people are riding camels at the beach	Two people are seated on a camel and another camel is in the foreground	۰/۴۲۵	۰/۴۵۸	۰/۹۶۹۷
	هیچ‌کسی در ساحل شترسواری نمی‌کند	دونفر روی یک شتر نشسته‌اند و شتر دیگری در تصویر دیده می‌شود			
۲	The man is standing in a lake near a waterfall	A man is performing a trick on a surfboard in the water	۰/۳۷۵	۰/۴۲۲۵	۰/۹۸۲۱
	مردی کنار دریاچه نزدیک آبشار ایستاده است	مردی روی تخته موج‌سواری تکنیک اجرا می‌کند			
۳	A person is reading the email	A person with a green shirt is jumping high over the grass	۰/۰۲۵	۰/۳۷۸	۰/۹۷۱۵
	شخصی دارد رایانامه خود را می‌خواند	شخصی با لباس سبز روی چمن می‌پرد			

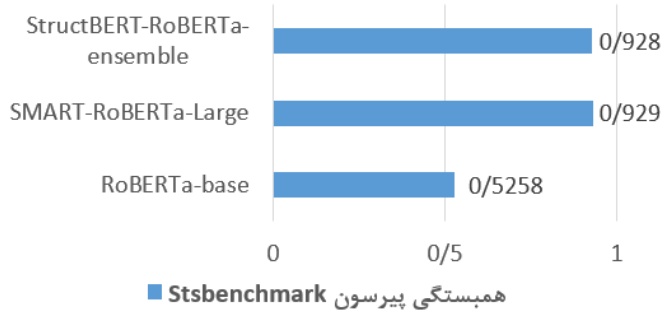


تصویر ۴-۱۹: مقایسه روش پیشنهادی VSD با روش‌های پیشین بهینه شده روی داده‌های SICK از دیدگاه همبستگی اسپیرمن

RoBERTa-base بهبود قابل توجهی داشته است. بنابراین همین عملکرد را در مورد بردارهای سایر توصیف‌گرها نیز می‌توان انتظار داشت. عملکرد مدل پایه[۲۹] و یا مدل بزرگ سیستم[۳۰] در ارتباط با این توصیف‌گر تفاوت چندانی نمی‌کند. این بدین معنی است که اضافه کردن پارامترهای بیشتر به سیستم از جمله اضافه کردن لایه‌های پنهان یا توجه (Attention heads) بیشتر، لزوما عملکرد سیستم را نسبت به حل مسئله خاص بهبود نمی‌دهد؛ نتایج نشان می‌دهد که بهینه‌سازی پارامترهای سیستم مدل زبانی استاندارد برای مسئله خاص بسیار مهمتر است. عملکرد روش پیشنهادی توصیف چندوجهی VSD در مقایسه با RoBerta-base روی داده‌های ارزیابی ربط SICK بهتر است، درحالی که این عملکرد روی داده‌های ارزیابی شباهت Stsbenchmark مشابه است.

[29] Base

[30] Large



تصویر ۴-۲۰: مقایسه روش RoBERTa با حالت‌های بهینه روی داده‌های stsbenchmark از دیدگاه همبستگی پیرسون

مزیت روش موجک

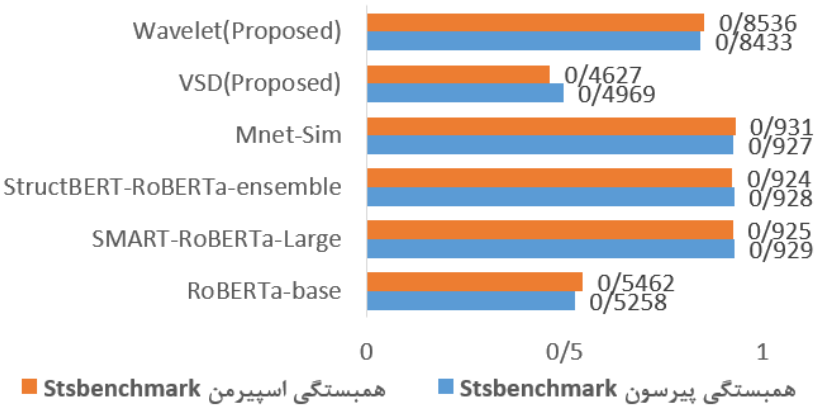
سیستم مبتنی بر موجک به دلیل تجزیه توصیف جمله به توصیف‌های جزئی‌تر می‌تواند نمایش بهتری نسبت به حالت عادی آن فراهم کند. تنها سیستمی که بهتر از روش موجک عمل کرده است روش SimCSE مبتنی بر مدل بزرگ‌تر RoBERTa است. به این معنا که تعداد ۲۴ لایه نورون‌های متصل به هم، در هرلایه ۱۰۲۴ نورون پنهان، ۱۶ واحد حافظه توجه و ۳۵۵ میلیون پارامتر استفاده کرده است. این در حالی است که مدل ما (به دلیل محدودیت سخت‌افزاری) مدل پایه بوده است که مدلی بسیار کوچکتر با ۱۲۵ میلیون پارامتر قابل تنظیم است؛ اما روش مورد استفاده قابلیت تعمیم به هرنوع سیستم توصیف‌گر جمله دارد. مقادیر بهینه برای رده‌بند توسط روش جستجوی GridSearch انتخاب شده است که از مجموعه‌ای پارامترها با توجه، به خطای رگرسیون، بهترین ترکیب را انتخاب می‌کند. تابع بهینه مورد استفاده تابع پایه شعاعی RBF می‌باشد. ضمن اینکه پارامتر بهینه جریمه تابع رگرسیون، مقدار ۰/۱ بود که نشان‌دهنده فاصله بالاتر تابع با نقاط داده نسبت به رگرسیون مورد استفاده روی داده‌های Stsbenchmark است.

تصویر ۴-۲۰ عملکرد RoBERTa را با نوع بهینه شده آن برای استنتاج متون مقایسه می‌کند. در تصویر مشخص است که روش‌های بهینه شده عملکرد بسیار بهتری دارند که نتایج تصویر ۴-۱۹ را نیز تأیید می‌کند.

دو روش SimCSE و توصیف‌گر چندوجهی (VSD) مستقیماً از معماری باناظر Siamese برای بهینه کردن بردار توصیف جمله‌ها استفاده می‌کنند. سایر روش‌ها از محاسبه میانگین مقادیر توصیف‌گر واژگان

برای تشکیل توصیف جمله استفاده می‌کنند [۳۱]. نتیجه بهتر SimCSE نسبت به تمامی روش‌ها خود گویای عملکرد بهتر معماری Siamese در تولید بردار توصیف‌گر است. تصویر ۴-۲۱ روش پیشنهادی VSD را در کنار برترین روش‌های اخیر روی داده‌های Stsbenchmark نمایش می‌دهد.

در تصویر ۴-۲۱ آمار بهترین روش‌ها روی مجموعه داده Stsbenchmark آمده است. لازم به ذکر است که مقادیر نوشته شده همگی در گزارش paperswithcode آمده است[۳۲].



تصویر ۴-۲۱: مقایسه روش پیشنهادی VSD با روش‌های پیشین بهینه شده روی داده‌های stsbenchmark از دو دیدگاه همبستگی پیرسون و اسپیرمن

با توجه به عملکرد سیستم VSD در مجموعه SICK انتظار می‌رفت که کارایی مشابهی روی Stsbenchmark داشته باشیم اما اینجا نتیجه ضعیف بوده است. از دلایل عدم موفقیت روش پیشنهادی VSD روی مجموعه داده شباهت را می‌توان موارد زیر برشمرد:

- ارزیابی روی داده‌های Stsbenchmark مشکل‌تر است به این دلیل که هدف آن ارزیابی شباهت است اما در داده‌های SICK نمره‌گذاری به منظور ارزیابی ارتباط بین دو جمله می‌باشد؛ و همانطور که می‌دانیم شباهت یک مورد خاص از ربط است و ارتباط دو جمله می‌تواند به شکل‌های مختلفی باشد از جمله شباهت، استلزام یا حتی تناقض.

- مجموعه داده شباهت، شامل داده‌های متنوع برگرفته شده از زیرنویس ویدیوها، اخبار از منابع مختلف، توضیح تصاویر، جمله‌های بازنویسی شده با عبارت‌های متفاوت و تالارهای پرسش و پاسخ

[31] Mean-pooling

[32] Papers with code (2021). Semantic Textual Similarity on SICK [online]. website: https://paperswithcode.com/sota/semantic-textual-similarity-on-sts-benchmark [accessed 15 July 2021

است و در آن تنوع داده و همچنین تنوع طول جمله‌های بیشتری وجود دارد.

- برخلاف مجموعه داده SICK در مجموعه داده شباهت Stsbenchmark، واژگان مخفف، تاکیدی و مبهم که نیاز به ابهام‌زدایی دارند بسیار دیده می‌شود، همچنین در مواردی تفاوت‌های دو جمله در مقادیر عددی مورد استفاده بین دو جمله است.

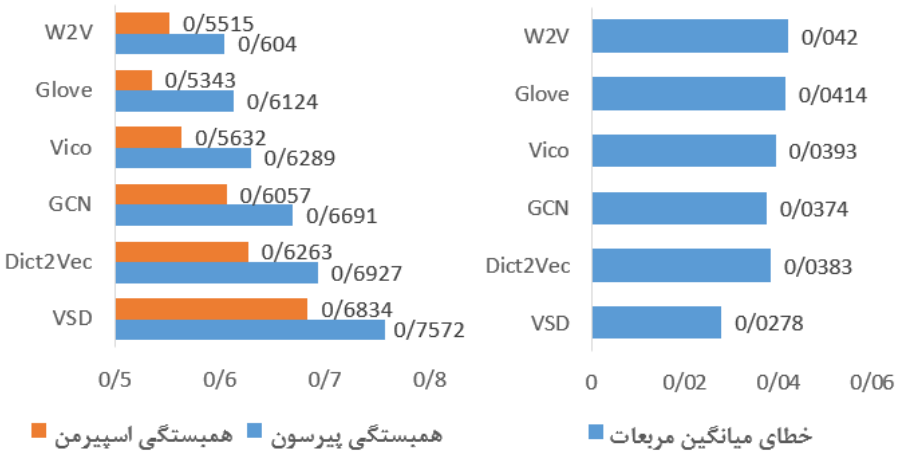
همانطور که می‌دانیم در دو روش VSD و StructBERT اطلاعات وابستگی واژگان جمله درون توصیف‌گر کدگذاری شده است. عملکرد روش StructBERT در حل مسئله ارزیابی شباهت زوج‌جمله‌ها بدون بهینه‌سازی دوباره وزن‌ها، درخور توجه است. به نظر می‌آید می‌توان فرض کرد که اطلاعات وابستگی وازگان با زویکرد الگوریتم StructBERT بسیار بهتر کدگذاری می‌شود.

بهترین عملکردها را سه روش StructBERT, SmartBERT و Mnet-sim کسب کردند. در لیست زیر به ویژگی‌های این سه روش که آن‌ها را از بقیه متمایز می‌کنند پرداختیم:

- شیوه آموزشی خصمانه که به شبکه عصبی اجازه نمی‌دهد تغییرات ناگهانی در وزن‌ها ایجاد شود و بدین ترتیب بیش‌برازش شبکه را تا حدود زیادی کنترل می‌کند. این روش اخیراً بسیار مورد توجه قرار گرفته چرا که با استفاده از این روش می‌توان سیستم‌های قدرتمندتری ساخت.

- اطلاعات وابستگی واژگان جمله به عنوان عناصر مهم در ارزیابی شباهت هستند، این مسئله در روش StructBERT به خوبی کدگذاری شده است چرا که عملکرد آن نسبت به RoBERTa معمولی بسیار بهتر بوده است.

- استفاده از اطلاعات معیارهای ارزیابی مانند فاصله کاسینوسی، اقلیدسی و غیره در قالب ویژگی‌هایی که به شبکه عمیق ارسال می‌شوند امری مهم تلقی می‌شود. نتیجه خوب سیستم پیشنهادی MNet در ارتباط با مجموعه SICK نیز بیانگر همین مسئله است.



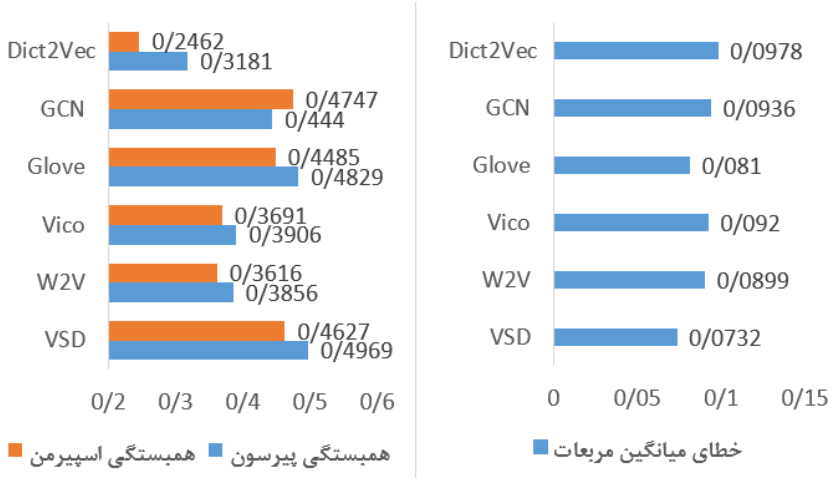
تصویر ۴-۲۲: نتایج شبکه عصبی توصیف‌گر چندوجهی روی داده‌های SICK

مزیت روش موجک

روش مبتنی بر موجک نسبت به سه روش برتر دیگر از مدل کوچکتری برای آموزش استفاده کرده است و با این حال توانسته نتایج موفقیت‌آمیزی کسب کند. روش موجک نسبت به رویکرد RoBERTa-base بهبود قابل توجهی داشته است. تبدیل موجک توانسته است تا اطلاعات طیفی محلی و زمانی را استخراج کند و اما در بخش رگرسیون بهترین نتیجه با تابع خطی SVR کسب شد که نشان‌دهنده روابط خطی بین ویژگی‌های مورد استفاده از تبدیل موجک برای محاسبه شباهت است. ضمن اینکه پارامتر بهینه جریمه خط رگرسیون مقدار ۱۰ بود که نشان‌دهنده فاصله کمتر تابع با نقاط داده نسبت به رگرسیون مورد استفاده روی داده‌های SICK است.

در تصویر ۴-۲۲ GCN اشاره به SynGCN دارد و VSD نماد روش پیشنهادی «توصیف‌گر چندوجهی» در بخش ۳-۱۴-۱ صفحه ۱۰۱ است. همان‌طور که در تصویر دیده می‌شود روش پیشنهادی VSD با توصیفگرهای چندگانه، از Glove و Word2Vec عملکرد بهتری داشته است. همچنین نتیجه آن در مقایسه با استفاده از تنها یک نوع توصیفگر بهتر است. می‌توان به این نتیجه‌گیری رسید که در این تصویر، بهترین مدل بین سه روش Dict2Vec, GCN و Vico از نوع توصیفگرهای Dict2Vec است که از واژگان همسایه مرتبط، در نقش کلیدی برای توصیف استفاده می‌کند.

در ارتباط با خطای میانگین مربعات، بهترین عملکرد را روش پیشنهادی توصیف‌گر چندوجهی VSD



تصویر ۴-۲۳: نتایج شبکه عصبی مبتنی توصیف‌گر چندوجهی روی داده‌های آزمون Stsbenchmark

داشته است. این بدین معنی می‌باشد که نمرات تولید شده توسط این سیستم، بیشترین شباهت را با نمره‌های انسانی دارد. این روش با اختلاف کمی از تمامی توصیف‌گرها بهتر عمکرده است در حالی که عملکرد Word2Vec از بقیه ضعیف‌تر بوده و سایر سیستم‌ها بین این دو قرار می‌گیرند اما در کل اختلاف قابل توجهی در این بخش وجود ندارد.

در تصویر ۴-۲۳، نتایج روش پیشنهادی VSD روی داده‌های ارزیابی شباهت Stsbenchmark آورده شده است؛ که نشان می‌دهد برخلاف نتایج خوب روش پیشنهادی روی داده‌های SICK، نتایج در ارزیابی شباهت Stsbenchmark ضعیف بوده است. ما پارامترهای بسیاری را در شبکه عصبی آزمودیم از جمله تعداد لایه‌های پنهان، روش کمینه‌سازی خطا و غیره، اما همبستگی بهتری به دست نیاوردیم. Dict2Vec از همه ضعیف‌تر عمل کرد اما VSD و GCN عملکرد بهتری از سایر روش‌ها داشته‌اند.

اما چرا؟ یک دلیل آن ممکن است این باشد که از وزن‌های پیش محاسبه شده شبکه عصبی برای آموزش شبکه استفاده نشده است. دیگر آنکه، داده‌های آموزش به تعداد کافی نیستند، هرچند که جمع‌آوری نمونه‌های بیشتر به دلیل نیاز به بازبینی داورهای انسانی امری مشکل است. دلیل دیگر آنکه شاید معماری LSTM به تنهایی برای حل این مسئله کافی نباشد. اگر مشکل فقط بهینه‌سازی بود انتظار می‌رفت نتایج با استفاده از AdaDelta, AdaGrad, RmsProp یا SGD تغییر کند اما نتیجه ضعیف‌تر و یا مشابه بود.

در جمع‌بندی می‌توان گفت که گرامر، نحوه قرارگیری واژگان و وابستگی به یکدیگر، به اندازه مقدار

جدول ۴-۲۶: مقایسه روش ما با روش‌های اخیر روی داده‌های آزمون SICK

روش	همبستگی پیرسون	همبستگی اسپیرمن	خطای میانگین مربعات
VSD	۰/۷۵۷۲	۰/۶۸۳۴	۰/۰۲۷۸
CNN-LSTM[۳۸]	۰/۸۶۳۴	۰/۷۹۹۱	۰/۲۶۵۴
n-gram attention[۷۹]	۰/۸۶۰	–	–
windowed self-attention[۵۵]	۰/۸۶۹۵	۰/۸۱۳۹	۰/۲۸۷۹
MultiTask[۲۷]	۰/۹۰۹	–	–
Multi-Embedding[۵۷]	۰/۸۸۷۶	–	–
Disan[۱۱۳]	۰/۸۶۹۵	۰/۸۱۳۹	۰/۲۸۷۹
LSTMs[۱۱۹]	۰/۸۶۷۶	۰/۸۰۸۳	۰/۲۵۳۲

جدول ۴-۲۷: مقایسه روش ما با روش‌های اخیر روی داده‌های آزمون Stsbenchmark

روش	همبستگی پیرسون
توصیف‌گر چندوجهی	۰/۴۹۶۹
ROT [۱۳۰]	۰/۸۰۶
n-gram attention [۷۹]	۰/۷۷۳
Unsupervised [۸]	۰/۶۷۲
MultiTask [۲۷]	۰/۸۰۳
Multi-Embedding [۵۷]	۰/۸۲۴۵

شباهت دو واژه اهمیت دارد. نکته مهم دیگر این است که هم‌رخدادی‌های موجود در تصاویر، می‌توانند اطلاعات مفید و جدیدی به خوبی هم‌رخدادی در متون فراهم کنند. از طرفی دیگر، استفاده از واژگان «تعریف واژه در فرهنگ لغات» برای توصیف، عملکرد مشابهی را مانند Glove و Word2Vec فراهم می‌کند.

در جداول ۴-۲۶ و ۴-۲۷ نتایج روش پیشنهادی VSD (توصیف‌گر چندوجهی) این‌بار با تعدادی از روش‌های پیشین که در فصل ۲ بررسی شده‌اند، مقایسه شده است. در مقایسه با نمودارهای قبل، این روش‌ها از توصیف‌گرهای واژگان مستقل از متن مانند Glove و Word2Vec استفاده کرده‌اند. متاسفانه، اگر در فرآیند آموزش، واژه‌ای باشد که توسط این دو توصیف‌گر دیده نشده است، برای آن برداری در نظر گرفته نمی‌شود.

ابتدا بیایید بررسی کنیم که دلیل موفقیت روش‌های برتر چیست؛ همان‌طور که در دو جدول ۴-۲۶ و ۴-۲۷ دیده می‌شود، به طور کلی، روش MultiTask بهترین نتایج را کسب کرده است. روشی که از

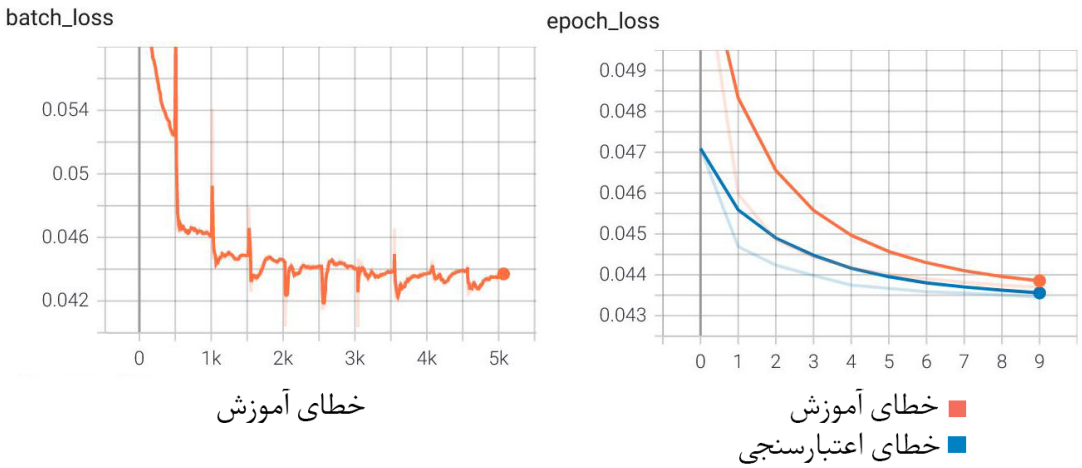
توصیفگرهای Word2Vec و فن توجه به واژگان (soft-alignment attention) استفاده می‌کند تا نحوه ارتباط بین اجزای دو جمله را کشف کند. نکته دیگر این روش، استفاده از داده‌های آموزشی بیشتر است. Disan تنها روش در لیست می‌باشد که مبتنی بر توجه مستقل به ویژگی‌های متفاوت توصیفگرها به صورت هم‌زمان است. پیشنهاد ما برای کارهای آینده استفاده از سایر توصیفگرها مانند SynGCN برای بهبود هردو روش یاد شده است. در جدول ۴-۲۶، موارد دقت وزنی در پنجره‌های همسایگی واژگان (Windowed self-attention) و دقت وزنی مبتنی بر چندنویسه (n-gram attention) تأیید می‌کنند که استفاده از فن توجه باعث بهبود قابل توجهی در نتایج همبستگی می‌شود. روش MultiEmbedding نیز نشان داده است که نتایج می‌تواند با استفاده از توصیفگرهای استاندارد هنوز بهبود یابد، به‌شرط آنکه (۱) معماری مورد استفاده برای تولید توصیفگر جمله نسبت به روش Siamese پیشرفته‌تر باشد یا (۲) بردارهای توصیف روی داده‌های آموزشی داده‌های ارزیابی شباهت از قبل بهینه شده باشند تا توصیف بهتری از واژگان فراهم کنند.

استفاده از معماری کدگذار-کدگشای CNN-LSTM توانسته است نتیجه خوبی با ارائه توصیف فشرده جمله، کسب کند. روش‌های بدون ناظر (Unsupervised) و روش ROT هردو از مدل‌های غیر عصبی در این جداول هستند که توانستند نتایج خوبی کسب کنند و در نتیجه می‌توانند بینش بهتری نسبت به روش‌های عصبی از نحوه ساخت توصیفگر مناسب جمله ارائه دهند.

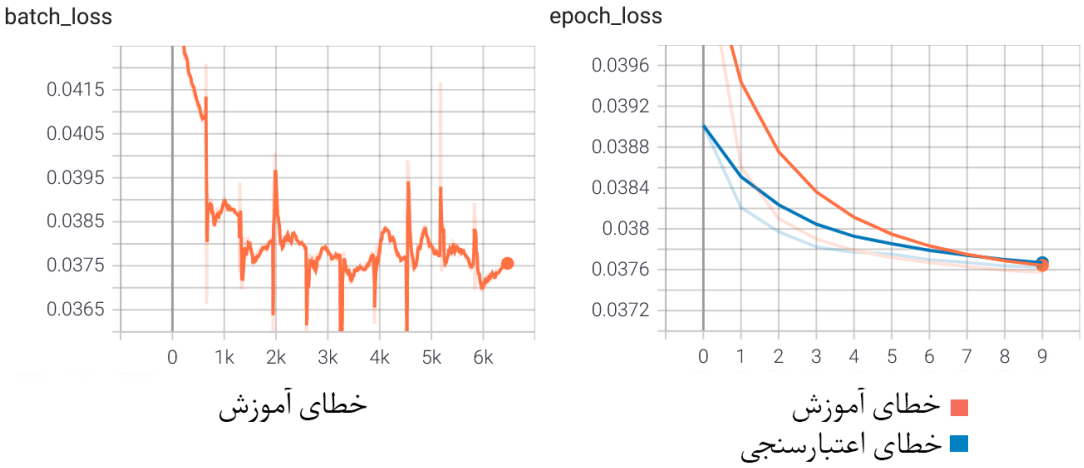
۴-۵-۱۰ ارزیابی شبکه عصبی کدگذار-کدگشا

در این بخش، سیستم طراحی شده در تصویر ۳-۵ بخش ۳-۱۴-۲ صفحه ۱۰۲ را ارزیابی می‌کنیم. تصاویر ۴-۲۴ و ۴-۲۵ فرآیند کاهش خطا با کمینه‌سازی میانگین حداقل مربعات در دو پایگاه داده SICK و Stsbenchmark نمایش می‌دهند.

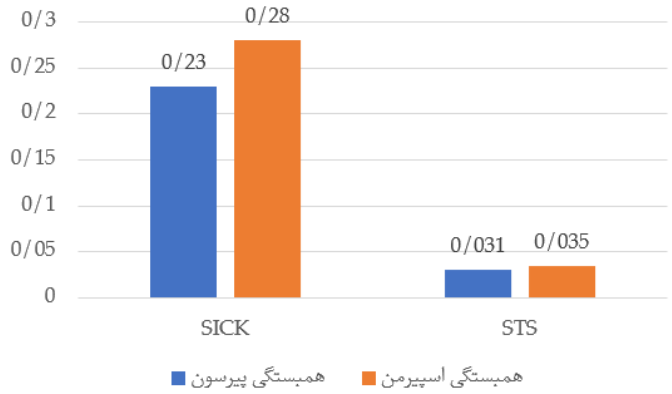
در دو تصویر دیده می‌شود هرچند که کمینه‌سازی به صورت پیوسته موفق نبوده است، اما این روند با استفاده از پنجره‌های پیچشی با وزن‌های مشترک کندتر بود و مقدار خطا بیشتر می‌باشد. کاهش خطا روی داده‌های Stsbenchmark موفق‌تر بوده است. با وجود کمتر نشدن خطای آموزش نسبت به



تصویر ۴-۲۴: روند کاهش خطا روی داده‌های SICK



تصویر ۴-۲۵: روند کاهش خطا روی داده‌های STS



تصویر ۴-۲۶: نتایج سیستم کدگذار-کدگشا در زمینه ارزیابی شباهت

اعتبارسنجی، آموزش بیشتر از ۷-۸ تکرار توجیه ندارد. حال به ارزیابی این روش توسط بردارهای تولید شده در بخش کدگذار، به‌منظور محاسبه شباهت بین دو جمله می‌پردازیم. بنابراین، ابتدا شبکه عصبی را توسط جمله‌های موجود در داده‌های بخش آموزشی هر پایگاه داده آموزش می‌دهیم، و برای آزمون، فاصله کاسینوسی بین دو بردار داده‌های آزمون را محاسبه می‌کنیم. اما نتایج این روش، همانطور که در تصویر ۴-۲۶ مشاهده می‌کنیم، در زمینه ارزیابی شباهت یا ربط ناموفق بوده است هرچند که در زمینه ارزیابی ربط عملکرد کمی بهتر می‌باشد.

۴-۶ بحث و بررسی

همانطور که در فصل ۲ بررسی شد، دو رویکرد کلی آماری و شبکه‌های عصبی برای حل مسئله ارزیابی شباهت زوج‌واژگان وجود دارد. دو مجموعه داده STSS65 و STSS131 قبل از داده‌های SICK و Stsbenchmark معرفی شدند. این دو پایگاه داده کوچک بودند و مناسب برای آموزش شبکه‌های عصبی نیستند. روش‌های آماری در این دو مجموعه داده، چه رویکردهای پیشین و چه آماری معرفی شده در فصل ۳، عملکرد خوبی داشتند اما تمامی این روش‌ها روی داده‌های SICK و Stsbenchmark عملکرد بسیار ضعیفی داشتند و مقادیر همبستگی کمتر از ۵۰ درصد کسب کردند. از دلایل این موضوع می‌توان به اشکال کلی در فرآیند هم‌ترازی واژگان بین دو جمله اشاره کرد یعنی جایی که اجزای مرتبط بین دو جمله با همدیگر تطبیق می‌یابند و شباهت بین اجزای هم‌تراز شده محاسبه می‌شود. عبارت‌ها در

این روش‌ها استفاده نشدند و ویژگی‌های استخراج شده از ماتریس شباهت زوج‌واژگان برای ارزیابی کافی نبودند. در این زمینه، کارهایی از جمله معرفی «ویژگی‌های محلی و عمومی در GLP» و همچنین تشکیل بردار توصیف‌گر واژگان مانند LSS و STASIS انجام شد و نتایج را کمی بهبود داد.

یک یافته بسیار جالب در آزمایش‌ها این بود که بردارهای توصیف‌گر واژگان در هنگام تعمیم به بردار جمله به صورت پیش‌فرض و استاندارد مؤثر نبودند. این موضوع برخلاف انتظار این پژوهش بود. اما در ادامه، بعد از بهینه کردن توصیف‌گرها روی «موضوع متن مجموعه داده‌های مورد ارزیابی»، عملکرد آن‌ها بهبود قابل توجهی یافت چنانچه این مسئله در تصویر ۴-۱۹ صفحه ۱۴۱ نیز مشخص شده است.

در یافته‌ای دیگر از این پژوهش، می‌توان با تلفیق روش‌های آماری با معیارهای مبتنی بر شبکه واژگان بهبود قابل توجهی در ارزیابی شباهت زوج‌فعل‌ها به دست آورد. این نتایج در جدول ۴-۱۹ صفحه ۱۲۳ نمایش داده شده است. همچنین این پژوهش تأیید می‌کند که افزودن اطلاعات آماری در ارزیابی شباهت می‌تواند بسیار مفید باشد.

یک یافته پیش‌بینی نشده عملکرد بهتر توصیف‌گرهای «مبتنی بر هم‌رخدادی واژگان در تصاویر ViCo» و «مبتنی بر درخت وابستگی اجزای جمله SynGCN» به صورت استاندارد، نسبت به روش Word2Vec است.

توصیف‌گرهای وابسته به محتوا روش‌های موفقی در زمینه توصیف جمله‌ها هستند، اگرچه عملکرد این روش‌ها با «بهینه کردن پارامترهای سیستم» برای «توصیف بهتر جمله‌ها» در «حل مسئله‌های خاص» امری لازم است. موفق‌ترین روش‌ها در زمینه توصیف جمله که برترین روش‌های حال حاضر در زمان نوشتن این رساله هستند، مبتنی بر توصیف‌گر BERT می‌باشند. این روش‌ها تلاش می‌کنند واژگان حذف شده از جمله را تخمین بزنند و بدین صورت توصیف خوبی از واژگان (و همچنین جمله‌ها) می‌سازند. در روش StructBERT افزودن اطلاعات وابستگی واژگان، به عبارتی ترتیب قرارگیری آن‌ها، باعث بهبود نتایج شده است. همچنین، در روش SimCSE روی داده‌های SICK مشاهده کردیم که استفاده از معماری Siamese نتایج را نسبت به حالت توصیف‌گرهای استاندارد بهبود می‌بخشد.

نتایج روش مبتنی بر تبدیل موجک روی توصیف‌گر جمله، نشان‌دهنده عملکرد خوب توابع موجک

روی بردار توصیف جمله می‌باشد. این بدان معنی است که «اطلاعات محلی سیگنال توصیف‌گر» برای ارزیابی شباهت مفید هستند. تعدادی دیگر از معیارهای فاصله غیر از نوع کاسینوسی از جمله فاصله مانهاتان و اقلیدسی آزموده شدند، اما یا عملکرد سیستم را مختل کردند یا بهبود قابل توجهی به سیستم اضافه نکردند و عملکرد معیار کاسینوسی در مقایسه، بهترین بود.

بعلاوه، یافته‌های این پژوهش می‌تواند به ما کمک کند تا با رویکردهای اخیر برای حل مسئله ارزیابی شباهت (زوج‌واژگان/زوج‌جمله‌ها) آشنا شویم. بهترین روش‌های پیشین، مبتنی بر شبکه‌های عمیق بودند و بردارهای توصیف‌گر واژگان به طور کلی عملکرد بسیار بهتری نسبت به روش‌های آماری دارند.

هنوز پژوهش‌های بسیاری باید صورت گیرد تا بشود توصیف‌گرها «بدون نیاز به بهینه‌سازی دوباره و به طور پیش‌فرض»، عملکرد قابل قبولی در ارزیابی شباهت یا استنتاج جمله داشته باشند. استفاده از فن توجه به «تمامی واژگان جمله به طور همزمان» نتایج بسیار بهتری نسبت به عدم وجود آن داشته است؛ طوری که پژوهش‌های اخیر همگی از این فن استفاده کرده‌اند و به نظر می‌اید وجود آن برای تشخیص عبارت‌ها و اصطلاح‌های زبانی مورد نیاز است.

۷-۴ جمع‌بندی

روش پیشنهادی WordnetPPDB روی داده‌های STSS65 در بیشتر آزمون‌ها موفق‌تر از روش‌های پیشین بوده است. رویکردهای دیگر پیشنهادی نیز نتایج قابل مقایسه‌ای با روش‌های دیگر دارند. اما با توجه به کوچک بودن پایگاه‌های داده STSS65 و STSS131، آزمون‌های بیشتری برای بررسی عملکرد روش‌های پیشنهادی موردنیاز است. نتایج روی داده‌های STSS131 بسیار مشابه با نتایج روش LSA بود. چهار روش پیشنهادی Relative, WordnetPPDB, Weighted و GLP توانسته‌اند نتایج بسیار خوبی روی هر دو پایگاه داده کسب کنند. نتایج روش شبکه عصبی نیز روی داده‌های بیشتر آزموده شده است که نشان‌دهنده «بهبود نتایج با استفاده از چندین توصیفگر واژگان به صورت هم‌زمان» نسبت به عملکرد مستقل آن‌ها می‌باشد. روش‌های مبتنی بر شبکه عصبی در حال حاضر بهترین روش‌ها هستند و استفاده از فن‌های توجه وزنی (Attention) بسیار در این شبکه‌ها رایج می‌باشد و توانسته نتایج را بهبود ببخشد.

توصیف‌گرهای وابسته به محتوا و مبتنی بر BERT که برای داده‌های ارزیابی شباهت بهینه شدند بهترین عملکرد را بین تمامی روش‌ها دارند. نتایج سیستم مبتنی بر تبدیل موجک نیز نشان‌دهنده «موفقیت این توابع در توصیف جزئی‌تر جمله» و «بهبود نتایج مدل زبانی پایه» است.

فصل ۵: نتیجه گیری

۵-۱ شباهت زوج‌واژگان

در این رساله، روشی ارائه شد که با افزودن اطلاعات آماری به ارزیابی شباهت زوج‌واژگان، توانست نتایج روش‌های پیشین را در پایگاه‌های داده STSS65 و STSS131 بهبود بخشد. مسئله مهم در حل شباهت توسط شبکه واژگان، امکان نبودن واژه‌های خاص در آن است، که با جایگزین کردن اطلاعات آماری، این مسئله مورد بررسی قرار گرفت. روش‌های مبتنی بر شبکه واژگان، عملکرد بهتری نسبت به نوع آماری مورد استفاده در این رساله داشتند، که دلیل آن ساختار درختی منظم و مشخص است تهیه شدهتوسط نیروی انسانی است. سیستم پیشنهادی «شباهت زوج‌واژگان به صورت تلفیقی» در زمینه ارزیابی زوج‌فعل‌ها بهتر از بقیه روش‌ها عمل کرد؛ از این جهت یافته‌های ما می‌تواند مکمل «پژوهش‌های پیشین در زمینه شبکه واژگان» تلقی شود. بزرگ‌ترین مشکل روش ما، ارزیابی واژگان نادر بود که می‌تواند موضوع پژوهش‌های آتی باشد. در نتیجه‌گیری از این بخش، به اهمیت در نظر گرفتن تمامی نقش‌واژه‌ها برای طراحی سیستم ارزیابی شباهت تأکید می‌شود.

۵-۲ شباهت جمله با تلفیق معیارهای آماری و شبکه واژگان

در این رساله بحث شد که روش‌های اخیر در زمینه «استخراج ویژگی» از «ماتریس شباهت زوج‌واژگان» بهینه عمل نکردند و بعلاوه، اهمیت محتوای اطلاعات یا به عبارتی بسامد رخداد واژگان نادیده گرفته شده است. دستاوردهای این بخش از رساله شامل موارد زیر می‌باشد:

- استفاده از YagoNet به عنوان شبکه واژگان به عنوان منبع مورد استفاده در محاسبه شباهت واژگان.
- استفاده از پایگاه داده جغرافیایی و ویکی‌پدیا برای شناسایی موجودیت‌ها در متن و اضافه کردن آن به نتایج ماتریس زوج‌واژگان.
- استفاده از اطلاعات آماری جمع‌آوری شده توسط گوگل برای محاسبه شباهت که در نتیجه آن،

در مواردی عملکرد سیستم بهبود یافت و در بعضی موارد، روی داده‌های ارزیابی ربط بهترین نتایج کسب شد.

- در روش GLP، اطلاعات محلی با پنجره‌های پیچشی از پنجره‌های شناور روی ماتریس زوج‌واژگان محاسبه شد و نتایج کمی بهبود یافت.

در این رساله ما تأیید می‌کنیم که استفاده از «معیار کوتاه‌ترین مسیر وزنی (صفحه ۳۷) در محاسبه شباهت توسط شبکه واژگان (Weighted Path)» و «پایگاه داده واژگان هم‌معنا (PPDB)» نتایج همبستگی را نسبت به روش‌های اخیر بهبود می‌دهد. در این رساله، استخراج ویژگی‌های مبتنی بر شبکه Yago در ماتریس زوج‌واژگان، همراه با شباهت برداری دو جمله، نتایج همبستگی را در داده‌های STSS65 بهبود بخشید. بهترین نتایج را در پایگاه داده STSS65 کسب کردیم در حالی که نتایج روی داده‌های STSS131 قابل مقایسه با روش آماری آنالیز معنای نهفته بود. در بسیاری از آزمون‌های کارآمدی غیر از همبستگی، چهار روش پیشنهادی Relative, WordnetPPDB, Weighted و GLP بهترین نتایج را کسب کردند. این مطالعه مکمل یافته‌های پژوهش هوآنگ و همکاران است [۵۴] که نشان داد استفاده از مقادیر «محتوای اطلاعات مفهوم مشترک بین دو واژه» ویژگی بسیار مهمی برای پیش‌بینی شباهت آن‌ها می‌باشد. بزرگ‌ترین مشکل روش‌های پیشنهادی (در این بخش) این بود که عبارت‌ها (چندواژه که فقط باهم معنی خاص می‌دهند) در نظر گرفته نشده بود؛ بنابراین پژوهش بیشتری در این بخش مورد نیاز است.

۵-۳ شباهت مبتنی بر شبکه عصبی

در این رساله اهمیت استفاده از توصیفگرهای واژگان در ارزیابی شباهت جمله‌ها بررسی شد. در این رساله مشخص شد که استفاده از توصیفگرهایی که مبتنی بر نحوه و ساختار جمله هستند اهمیت یکسانی در مقابل استفاده از توصیفگرهای معنایی دارد. شواهد حاصل از این پژوهش نشان می‌دهد که استفاده از فن توجه به واژگان (Attention)، هم‌زمان با استفاده از چند توصیفگر متنوع، می‌تواند عملکرد سیستم

ارزیابی شباهت را بهبود دهد؛ بنابراین، یافته‌های این رساله مکمل نتایج پژوهش‌های گذشته در زمینه ارزیابی توصیف‌گرهای SynGCN, Dict2Vec و ViCo است [۱۲۸، ۱۲۲، ۴۱]. این درک جدید می‌تواند به بهبود پیش‌بینی‌های مشابهت متن و سیستم‌های وابسته کمک کند. در این مطالعه استفاده از نوع سازوکار فن توجه ارزیابی نشد زیرا هدف این تحقیق نبوده است. یافته‌های این رساله دارای چندین پیامد مهم برای پژوهش‌های آینده می‌باشد، زیرا یافته‌ها تأکید می‌کند که نیاز به پژوهش بیشتری در جهت طراحی توصیفگرهای مبتنی بر نحو وجود دارد و اینکه نباید به تنهایی از توصیفگرهای Word2Vec و Glove برای طراحی سیستم‌های آینده استفاده کرد.

در ادامه این رساله، توصیف‌گرهای وابسته به محتوا نیز ارزیابی شدند. این توصیف‌گرها برای حل مشکل ابهام‌زدایی از معنای واژگان، «وابسته به محتوایی که واژه در آن رخ می‌دهد»، بردار توصیف‌گر متفاوتی تولید می‌کنند. با وجود طراحی مدل‌های زبانی قوی از جمله RoBERTa، مشکل این روش‌ها در حالت بهینه نشده، این است که عملکردشان در زمینه حل مسائل خاص از جمله ارزیابی شباهت زوج‌جمله‌ها، ضعیف‌تر از حد انتظار بوده است. در نمودارها دیدیم که این عملکرد با بهینه‌سازی توصیف‌گرهای واژگان یا جمله‌ها (بسته به رویکرد) روی داده‌های آموزش، نتایج بسیار بهبود یافت. برترین مدل‌های کنونی این نوع مدل‌ها هستند که بهترین نتایج را نسبت به سایر روش‌های بررسی شده روی مجموعه داده‌های SICK و Stsbenchmark کسب کردند.

نتیجه‌گیری روش موجک

در روش پیشنهادی مبتنی بر توصیف موجک، از توصیف‌گر جمله که میانگین توصیف واژگان تشکیل دهنده است (می‌تواند مثلا آخرین خروجی لایه پنهان شبکه هم باشد) تبدیل گرفته شده است. برای کارهای آینده، پیشنهاد این است که تبدیل موجک روی توصیف مستقل واژگان گرفته شود تا شاید باعث بهبود بیشتر کارایی سیستم شود. به طور کلی رویکرد موجک موفقیت‌آمیز بوده است و می‌توان در این زمینه، پژوهش‌های بیشتری انجام داد.

در این پژوهش مشخص شد که روش‌های آماری تلاش می‌کنند بهترین هم‌ترازی بین واژه‌های دو

جمله را بیابند. بعضی روش‌های عصبی با استفاده از «فن توجه» واژگان کلیدی جمله‌ها را برای تولید توصیف‌گر شناسایی می‌کنند و اینگونه، توانایی به خاطر سپردن اطلاعات انتخابی و مهم را به شبکه عصبی اضافه می‌کنند. به نظر می‌آید نیاز به داده‌های بیشتر و طراحی مدل‌های بدون ناظر وجود دارد که بتوانند توصیف خوبی از جمله را با توجه به واژگان تشکیل‌دهنده آن بسازند. با وجود اینکه در این پژوهش، هدف، ارزیابی شباهت زوج‌جمله‌ها بوده است، یافته‌های آن می‌تواند در زمینه‌های درک زبان طبیعی، خلاصه‌سازی متون و سیستم‌های پرسش و پاسخ نیز استفاده شود. با وجود محدودیت‌های موجود در این پژوهش، یافته‌ها به درک بهتر مسیر پژوهش‌های پیشین و آینده کمک می‌کند. همچنین پژوهش‌های بیشتری برای ارزیابی اهمیت «فن توجه به واژگان» در شباهت جمله‌ها مورد نیاز است.

۵-۴ پژوهش‌های آینده

برای پژوهش‌های آتی توصیه می‌شود که روش‌های یادگیری تقویتی به عنوان جایگزین بهینه‌ساز پارامترها، در طراحی سیستم رگرسیون استفاده شود. سیستم یادگیری تقویتی می‌تواند با توجه به اختلاف مقدار شباهت هدف و مقدار پیش‌بینی‌شده، بعضی نمونه‌ها را بیشتر جریمه کند. همچنین اهمیت وزن‌دهی نمونه‌ها در محاسبه پارامترهای بهینه سیستم، می‌تواند افزوده شود.

در سیستم‌های مبتنی بر BERT واژگان به صورت تصادفی ماسک می‌شوند؛ مسئله‌ای که وجود دارد این است که گاهی نام موجودیت‌ها دارای ابهام نیست اما رویه تولید توصیف‌گر برای تمام موجودیت‌ها یکسان می‌باشد و در نتیجه موجودیت‌های بین دو جمله ممکن است اندازه شباهت مشابهی داشته باشند در حالی که واقعیت اینطور نیست. بنابراین یک ساز و کار مناسب برای این موضوع باید پیشنهاد شود. روش‌های آموزش شبکه‌های عصبی هم با وجود افزایش پارامترها هنوز نیازمند تغییرات از جمله موازی کردن آموزش است.

استفاده از مبدل‌ها[1] جایگزین موفقی برای LSTM بوده است اما مشکل «پیچیدگی محاسباتی بالا» و «نیاز به حافظه بالا» استفاده از آن‌ها را مشکل کرده است. همچنین عملکرد «غیرخطی مدل‌های زبانی

[1] Transformers

پیش‌آموزش دیده»، درک نحوه تصمیم‌گیری آن‌ها را مشکل می‌کند و نیاز به پژوهش برای درک بیشتر در این زمینه وجود دارد.

از روش‌های آموزش دیگر که در زمینه بینایی ماشین، موفقیت‌آمیز مورد استفاده قرار گرفته است، راهبرد هجوم خصمانه[۲] می‌باشد. در این شیوه، به مدل برای حل مسئله‌ای خاص مانند مساوی بودن معنی دو جمله، متونی داده می‌شود که غلط‌انداز باشد ولی برای انسان قابل تشخیص است. در این رویکرد، دو مدل وجود دارد که در آن یکی از آن‌ها (مدل خصمانه) سعی می‌کند با تولید جمله‌های غلط‌انداز، توصیف‌گر مدل زبانی را به اشتباه بیندازد و مدل زبانی تلاش می‌کند تا تشخیصش درست باشد. مدل خصمانه به منظور فریب بهتر مدل زبانی، و مدل زبانی به منظور تهیه توصیف مناسب‌تر از جمله بهینه می‌شوند. از چالش‌های تولید خودکار متن غلط‌انداز، این است که باید از نظر گرامری و معنایی درست باشد. در کل، تولید این‌گونه متن‌ها توسط مدل خصمانه به صورت خودکار مشکل است، هرچند که این شیوه آموزشی می‌تواند در کارایی مدل‌های زبانی در حل مسائل خاص بهبود قابل توجهی ایجاد کند [۱۰۵].

[2] Adverserial attacks

مراجع

[1] Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. E., and Arshad, H. (2018). State-of-the-art in Artificial Neural Network Applications: a Survey.

[2] Achananuparp, P., Hu, X., Zhou, X., and Zhang, X. (2008). Utilizing sentence similarity and question type similarity to response to similar questions in knowledge-sharing community. *Proceedings of QAWeb Workshop*, 214.

[3] Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A Study on Similarity and Relatedness using Distributional and Wordnet-based Approaches. In *NAACL HLT - Human Language Technologies: the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

[4] Ahlgren, P. and Colliander, C. (2009). Document-document Similarity Approaches and Science Mapping: Experimental Comparison of Five Approaches. *Journal of informetrics*, 3(1).

[5] Aji, S. and Kaimal, R. (2012). Document Summarization using Positive Pointwise Mutual Information. *International Journal of Computer Science and Information Technology*.

[6] Aletras, N. and Stevenson, M. (2015). A Hybrid Distributional and Knowledge-based

Model of Lexical Semantics. In *Proceedings of the 4th Joint Conference on Lexical and Computational Semantics*, pages 20–29.

[7] Ali, Z. (2019). A Simple Word2vec Tutorial.

[8] Arroyo-Fernández, I., Méndez-Cruz, C. F., Sierra, G., Torres-Moreno, J. M., and Sidorov, G. (2019). Unsupervised Sentence Representations as Word Information Series: Revisiting TF–IDF. *Computer Speech and Language*, 56:107–129.

[9] Baeza-Yates, R., Ribeiro-neto, B., Mills, D., Bonn, o., Juan, S., Mexico, M., Taipei, C., Wesley, A., and Limited, L. (1999). *Modern Information Retrieval*. Addison-Wesley.

[10] Baker, S., Reichart, R., and Korhonen, A. (2014). An Unsupervised Model for instance Level Subcategorization Acquisition. *EMNLP - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 278–289.

[11] Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't Count, Predict! a Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL*, 1:238–247.

[12] Bendersky, E. (2015). Memory layout of multi-dimensional arrays.

[13] Berndt, D. and Clifford, J. (1994). using Dynamic Time Warping To Find Patterns in Time Series. In *Workshop on Knowledge Knowledge Discovery in Databases*, volume 398, pages 359–370.

[14] Bilgin, M. and Şenturk, I. F. (2017). Sentiment analysis on Twitter data with semi-supervised Doc2Vec. In *2nd International Conference on Computer Science and Engineering, UBMK 2017*.

[15] Bock, T. (2020). Singular Value Decomposition (Svd) Tutorial using Examples in R.

[16] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *CoRR*, abs/1607.0.

[17] Bolshakov, I. A. and Gelbukh, A. (2004). *Computational Linguistics: Models, Resources, Applications*. Instituto Politécnico Nacional IPN-UNAM-FCE Publication.

[18] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A Large Annotated Corpus for Learning Natural Language inference. In *Conference Proceedings - Emnlp 2015: Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

[19] Brants, T. and Franz, A. (2006). Web 1t 5-gram Corpus Version 1.1. *Linguistic Data Consortium*.

[20] Brown, P. E., Pietra, V. J. D., Mercer, R. L., Pietra, S. a. D., and Lai, J. C. (1992). An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics*, 18(1):31–40.

[21] Bruni, E., Uijlings, J., Baroni, M., and Sebe, N. (2012). Distributional Semantics with Eyes: using Image Analysis To Improve Computational Representations of Word Meaning. *Proceedings of the 20Th ACM International Conference on Multimedia*, pages 1219–1228.

[22] Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes*, 25(2-3).

[23] Buscaldi, D., Rosso, P., Gómez-Soriano, J. M., and Sanchis, E. (2010). Answering

Questions with an N-Gram based Passage Retrieval Engine. *Journal of Intelligent Information Systems*, 34(2):113–134.

[24] Carullo, M., Binaghi, E., and Gallo, I. (2009). An online Document Clustering Technique for Short Web Contents. *Pattern Recognition Letters*, 30(10).

[25] Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, i., and Specia, L. (2018). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of SemEval*.

[26] Chen, D. L. and Dolan, W. B. (2011). Collecting Highly Parallel Data for Paraphrase Evaluation. In *ACL-HLT - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA.

[27] Choi, H. S. and Lee, H. (2019). Multitask Learning Approach for Understanding the Relationship Between Two Sentences. *Information Sciences*, 485:413–426.

[28] Consortium, T. B. (2007). the British National Corpus, Version 3.

[29] Croft, D., Coupland, S., Shell, J., and Brown, S. (2013). A Fast and Efficient Semantic Short Text Similarity Metric. In *13Th UK Workshop on Computational Intelligence*, pages 221–227.

[30] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

[31] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Corr*, abs/1810.0.

[32] Ene, A. and Wayne, K. (2006). COS 226 WordNet.

[33] Everingham, M., Eslami, S. M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, a. (2014). the Pascal Visual Object Classes Challenge: a Retrospective. *International Journal of Computer Vision*, 111:98–136.

[34] Feng, G., Li, S., Sun, T., and Zhang, B. (2018). A Probabilistic Model Derived Term Weighting Scheme for Text Classification. *Pattern Recognition Letters*, 110.

[35] Feng, J., Zhou, Y., and Martin, T. (2008). Sentence Similarity based on Relevance. *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU*, pages 832–839.

[36] Ferrández, A., Maté, A., Peral, J., Trujillo, J., De Gregorio, E., and Aufaure, M. A. (2016). A Framework for Enriching Data Warehouse Analysis with Question Answering Systems. *Journal of Intelligent Information Systems*, 46(1).

[37] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing Search in Context: the Concept Revisited. *Proceedings of the 10Th International Conference on World Wide Web*, 20(1):406–414.

[38] Fu, Q., Wang, C., and Han, X. (2020). A CNN-LSTM Network with Attention Approach for Learning Universal Sentence Representation in Embedded System. *Microprocessors and Microsystems*, 74:103051.

[39] Gao, T., Yao, X., and Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv*.

[40] Goyal, P., Pandey, S., and Jain, K. (2018). *Deep Learning for Natural Language Processing: Creating Neural Networks with Python*. Apress.

[41] Gupta, T., Schwing, A., and Hoiem, D. (2019). ViCo: Word Embeddings from Visual Co-Occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea.

[42] Hadj Taieb, M. A., Ben Aouicha, M., and Ben Hamadou, A. (2013). Computing Semantic Relatedness using Wikipedia Features. *Knowledge-based Systems*, 50:260–278.

[43] Hahn, U. (2014). Similarity. *Wiley interdisciplinary Reviews: Cognitive Science*.

[44] Halawi, G., Dror, G., Gabrilovich, E., and Koren, Y. (2012). Large-Scale Learning of Word Relatedness with Constraints. In *Proceedings of the 25Th ACM Sigkdd International Conference on Knowledge Discovery*

Data Mining, number September, pages 1406–1414.

[45] Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical Correlation Analysis: an Overview with Application to Learning Methods.

[46] Harris, Z. S. (1954). Distributional Structure. *Word*.

[47] Hauke, J. and Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, 30:87–93.

[48] Haveliwala, T., Kamvar, S., and Jeh, G. (2003). An Analytical Comparison of Approaches To Personalizing Pagerank. Technical report, Stanford University.

[49] Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

[50] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9:1735–1780.

[51] Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47(1):853–899.

[52] Hu, X., Zhang, X., Lu, C., Park, E. K., and Zhou, X. (2009). Exploiting Wikipedia as External Knowledge for Document Clustering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[53] Huang, E. H., Socher, R., Manning, C. D., and Ng, a. Y. (2012). Improving Word Representations via Global Context and Multipleword Prototypes. In *50Th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*.

[54] Huang, H., Wu, H., Wei, X., Gao, Y., and Shi, S. (2019). Mapping Sentences to Concept Transferred Space for Semantic Textual Similarity. *Knowledge and Information Systems*, 60(3):1353–1376.

[55] Huang, T., Deng, Z. H., Shen, G., and Chen, X. (2020). A Window-based Self-Attention Approach for Sentence Encoding. *Neurocomputing*, 375(January):25–31.

[56] Huddleston, R. and Pullum, G. K. (2002). *the Cambridge Grammar of the English Language*. Cambridge University Press.

[57] Huy, N., Le, N. M., Tomohiro, Y., and Tatsuya, I. (2019). Sentence Modeling via Multiple Word Embeddings and Multi-Level Comparison for Semantic Textual Similarity. *Information Processing and Management*, 56(6):102090.

[58] Imam, A. (2020). Neural Networks (Part 1).

[59] Islam, A. and inkpen, D. (2008). Semantic Text Similarity using Corpus-based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2):1–25.

[60] Islam, A., Milios, E., and Kešelj, V. (2012). Text Similarity using Google Tri-Grams. In *Canadian Conference on Artificial Intelligence*, pages 312–317, Toronto, Canada.

[61] Islam, M. A. and inkpen, D. (2006a). Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC*, pages 1033–1038, Genoa, Italy.

[62] Islam, M. A. and inkpen, D. (2006b). Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words. In *the Fifth International Conference on Language Resources and Evaluation*.

[63] James, G., Witten, D., and Hastie, T. (2019). *Introduction To Statistical Learning with Applications in R*. Springer.

[64] Jeyaraj, M. N. and Kasthurirathna, D. (2021). MNet-Sim : a Multi-layered Semantic Similarity Network to Evaluate Sentence Similarity. *International Journal of Engineering Trends and Technology*, 69(7):181–189.

[65] Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Zhao, T. (2020). SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. *arXiv*.

[66] Jimenez, S., Gonzalez, F. A., Gelbukh, A., and Duenas, G. (2019). Word2Set: Wordnet-based Word Representation Rivaling Neural Word Embedding for Lexical Similarity and Sentiment Analysis. *IEEE Computational Intelligence Magazine*, 14(2):41–53.

[67] Kartikeya, R. (2020). Pooling Layer - Short and Simple. Here's All the Information You Should Know.

[68] Kilgarriff, A. and Fellbaum, C. (2000). WordNet: an Electronic Lexical Database. *Language*, 76(3).

[69] Kotlerman, L., Dagan, I., and Kurland, O. (2018). Clustering Small-sized Collections of Short Texts. *Information Retrieval Journal*, 21(4).

[70] Kumar Tomar, N. (2021). Convolution Neural Network (CNN) – Fundamental of Deep Learning.

[71] Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

[72] Lan, W. and Xu, W. (2018). Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language inference, and Question Answering . In *27th International Conference on Computational Linguistics*, pages 3890–3902, Santa Fe, New Mexico, USA.

[73] Larkey, L. B. and Markman, A. B. (2005). Processes of Similarity Judgment. *Cognitive Science*, 29(6):1061–1076.

[74] Li, C., Xu, B., Wu, G., Zhuang, T., Wang, X., and Ge, W. (2014). Improving Word Embeddings via Combining with Complementary Languages. In *Lecture Notes*

in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

[75] Li, Y., Bandar, Z. A., and McLean, D. (2003). An Approach for Measuring Semantic Similarity Between Words using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882.

[76] Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., and Crockett, K. (2006). Sentence Similarity based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150.

[77] Liu, X., Zhou, Y., and Zheng, R. (2007). Sentence Similarity based on Dynamic Time Warping. In *Icsc 2007 International Conference on Semantic Computing*, pages 250–256. IEEE.

[78] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: a Robustly Optimized BERT Pretraining Approach. *arXiv*.

[79] Lopez-Gazpio, I., Maritxalar, M., Lapata, M., and Agirre, E. (2019). Word n-gram Attention Models for Sentence Similarity and inference. *Expert Systems with Applications*, 132:1–11.

[80] Luong, M. T., Socher, R., and Manning, C. D. (2013). Better Word Representations with Recursive Neural Networks for Morphology. In *Conll - 17Th Conference on Computational Natural Language Learning*.

[81] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). the Stanford Core NLP Natural Language Processing Toolkit. In *Proceedings*

of 52Nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

[82] Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK Cure for the Evaluation of Compositional Distributional Semantic Models. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC*, pages 216–223, Reykjavik, Iceland.

[83] McCrae, J. P., Rudnicka, E., and Bond, F. (2019). English WordNet: a New Open-source Wordnet for English.

[84] Mijangos, V., Sierra, G., and Montes, A. (2017). Sentence Level Matrix Representation for Document Spectral Clustering. *Pattern Recognition Letters*, 85.

[85] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

[86] Mikolov, T., Chen, K., S Corrado, G., and A. Dean, J. (2013b). Computing Numeric Representations of Words in a High-dimensional Space.

[87] Miller, G. A. (1995). WordNet: a Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

[88] Miller, G. A. and Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*.

[89] Mohamed, I. S. (2020). *Detection and Tracking of Pallets using a Laser Rangefinder and Machine Learning Techniques*. PhD thesis, University of Genova, Italy.

[90] Mueller, J. and Thyagarajan, A. (2016). Siamese Recurrent Architectures for Learning Sentence Similarity. In *30th AAAI Conference on Artificial Intelligence*, pages 2786–2792, Phoenix, Arizona, USA.

[91] Needleman, S. B. and Wunsch, C. D. (1970). A General Method Applicable To the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, 48(3):443–453.

[92] Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (2004). The University of South Florida Free Association, Rhyme, and Word Fragment Norms.

[93] Ni, Y., Xu, Q. K., Cao, F., Mass, Y., Sheinwald, D., Zhu, H. J., and Cao, S. S. (2016). Semantic Documents Relatedness using Concept Graph Representation. In *WSDM - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*.

[94] Niwa, Y. and Nitta, Y. (1994). Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries. In *COLinG : Proceedings of the 15th conference on Computational linguistics*, pages 304–309.

[95] Oliva, J., Serrano, J. I., Del Castillo, M. D., and Iglesias, Á. (2011). SyMSS: a Syntax-based Measure for Short-text Semantic Similarity. *Data and Knowledge Engineering*, 70(4):390–405.

[96] O'Shea, J., Bandar, Z., Crockett, K., and McLean, D. (2008). A Comparative Study of Two Short Text Semantic Similarity Measures. In *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4953 LNAI, pages 172–181.

[97] Pagliardini, M., Gupta, P., and Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings using Compositional N-Gram Features. In *NAACL HLT - Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 528–540, New Orleans, Louisiana, USA.

[98] Patwardhan, S. and Pedersen, T. (2006). Using Wordnet-based Context Vectors To Estimate the Semantic Relatedness of Concepts. In *In 11Th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

[99] Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2015). PPDB 2.0: Better Paraphrase Ranking, Fine-Grained Entailment Relations, Word Embeddings, and Style Classification. In *ACL-Ijcnlp 2015 - 53Rd Annual Meeting of the Association for Computational Linguistics and the 7Th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, volume 2, pages 425–430, Beijing, China. Association for Computational Linguistics.

[100] Pawar, A. and Mago, V. (2018). Calculating the Similarity Between Words and Sentences using a Lexical Database and Corpus Statistics. *arXiv*, pages 1–14.

[101] Pennington, J., Socher, R., and Manning, C. D. (2014a). Glove: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (Emnlp)*, pages 1532–1543.

[102] Pennington, J., Socher, R., and Manning, C. D. (2014b). GloVe: Global Vectors

for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1532–1543.

[103] Phi, M. (2018). Illustrated Guide to LSTM's and GRU's: a Step by Step Explanation.

[104] Prakoso, D. W., Abdi, A., and Amrit, C. (2021). Short Text Similarity Measurement Methods: a Review. *Soft Computing*.

[105] Qiu, X. P., Sun, T. X., Xu, Y. G., Shao, Y. F., Dai, N., and Huang, X. J. (2020). Pre-trained Models for Natural Language Processing: a Survey. *Science China Technological Sciences*, 63(10):1872–1897.

[106] Quirk, R. (1962). *the Use of English*. Longman, London, 2nd edition.

[107] Rakib, M. R. H., Islam, A., and Milios, E. (2016). Phrase Relatedness Function using Overlapping Bi-Gram Context. In *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

[108] Rakib, R. H., Islam, A., and Milios, E. (2018). Improving Text Relatedness by Incorporating Phrase Relatedness with Word Relatedness. *Computational Intelligence*, 34(3):939–966.

[109] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv*.

[110] Rubenstein, H. and Goodenough, J. B. (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

[111] Sahishanu (2020). LSTM – Derivation of Back Propagation Through Time.

[112] Shen, G., Deng, Z. H., Huang, T., and Chen, X. (2020). Learning to Compose Over Tree Structures via Pos Tags for Sentence Representation. *Expert Systems with Applications*, 141.

[113] Shen, T., Jiang, J., Zhou, T., Pan, S., Long, G., and Zhang, C. (2018). Disan: Directional Self-attention Network for RNN/CNN-free Language Understanding. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 5446–5455.

[114] Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., and Rezende, S. O. (2019). Knowledge-Enhanced Document Embeddings for Text Classification. *Knowledge-based Systems*, 163:955–971.

[115] Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: a Large ontology from Wikipedia and Wordnet. *Web Semantics*.

[116] Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision*.

[117] Sutskever, I., Martens, J., and Hinton, G. (2011). Generating Text with Recurrent Neural Networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1017–1024, Madison, WI, USA. Omnipress.

[118] Szumlanski, S., Gomez, F., and Sims, V. K. (2013). A New Set of Norms for Semantic Relatedness Measures. *ACL - 51St Annual Meeting of the Association for Computational Linguistics*, 2:890–895.

[119] Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks. In *ACL-IJCNLP - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, pages 1556–1566, Beijing, China.

[120] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the Gap To Human-Level Performance in Face Verification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[121] Tien, N. H., Le, N. M., Tomohiro, Y., and Tatsuya, I. (2019). Sentence Modeling via Multiple Word Embeddings and Multi-Level Comparison for Semantic Textual Similarity. *Information Processing and Management*, 56(6).

[122] Tissier, J., Gravier, C., and Habrard, A. (2017). Dict2vec : Learning Word Embeddings using Lexical Dictionaries. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 254–263.

[123] Tolosana, R., Vera-RodrÍguez, R., Fierrez, J., and Ortega-Garcia, J. (2018). Exploring Recurrent Neural Networks for on-Line Handwritten Signature Biometrics. *IEEE Access*, 6:5128–5138.

[124] Tsatsaronis, G., Varlamis, I., and Vazirgiannis, M. (2010). Text Relatedness based on a Word thesaurus. *Journal of Artificial Intelligence Research*, 37:1–39.

[125] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, LIX(236):433–460.

[126] Vanderwende, L. and Dolan, W. B. (2006). What Syntax Can Contribute in the Entailment Task. In *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3944 LNAI.

[127] Vani, K. and Gupta, D. (2017). Text Plagiarism Classification using Syntax based Linguistic Features. *Expert Systems with Applications*, 88.

[128] Vashishth, S., Bhandari, M., Yadav, P., Rai, P., Bhattacharyya, C., and Talukdar, P. (2019). Incorporating Syntactic and Semantic Information in Word Embeddings using Graph Convolutional Networks. In *Proceedings of the 57Th Annual Meeting of the Association for Computational Linguistics*, pages 3308–3318, Florence, Italy. Association for Computational Linguistics.

[129] Wang, W., Bi, B., Yan, M., Wu, C., Bao, Z., Xia, J., Peng, L., and Si, L. (2019). StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding.

[130] Wang, Z., Zhang, Y., and Wu, H. (2020). Structural-Aware Sentence Similarity with Recursive Optimal Transport. *Arxiv*.

[131] Wei, T., Lu, Y., Chang, H., Zhou, Q., and Bao, X. (2015). A Semantic Approach for Text Clustering using Wordnet and Lexical Chains. *Expert Systems with Applications*, 42(4):2264–2275.

[132] Wen, C. (2021). Comparison of insurance Cost Prediction with and Without Hidden Layer in Neural Networks.

[133] Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016). Charagram: Embedding Words and Sentences via Character n-grams. In *EMNLP - Conference on Empirical Methods in Natural Language Processing*.

[134] Wikipedia (2021a). Perplexity.

[135] Wikipedia (2021b). Spearman's Rank Correlation Coefficient.

[136] Wu, H. and Huang, H. (2016). Sentence Similarity Computational Model based on Information Content. *IEICE Transactions on Information and Systems*, E99D(6):1645–1652.

[137] Wu, Z. and Palmer, M. (1994). Verbs Semantics and Lexical Selection. In *ACL : Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico.

[138] Yamada, I., Asai, A., Shindo, H., Takeda, H., and Takefuji, Y. (2018). Wikipedia2Vec: an Optimized Tool for Learning Embeddings of Words and Entities from Wikipedia. *CoRR*.

[139] Yang, D. and Powers, D. M. (2005). Verb Similarity on the Taxonomy of Wordnet. *GWC: 3Rd International Global Wordnet Conference*, pages 121–128.

[140] Yang, J., Li, Y., Gao, C., and Zhang, Y. (2021). Measuring the Short Text Similarity based on Semantic and Syntactic Information. *Future Generation Computer Systems*, 114:169–180.

[141] Yi, D., Lei, Z., and Li, S. (2014). Deep Metric Learning for Practical Person Re-Identification. *Proceedings - International Conference on Pattern Recognition*.

[142] Zhang, D., Peng, Q., Lin, J., Wang, D., Liu, X., and Zhuang, J. (2019). Simulating Reservoir Operation using a Recurrent Neural Network Algorithm. *Water*, 11:865.

[143] Zhang, H. and Chow, T. W. (2012). A Multi-level Matching Method with Hybrid Similarity for Document Retrieval. *Expert Systems with Applications*, 39(3).

[144] Zheng, W., Liu, X., and Yin, L. (2021). Sentence Representation Method based on Multi-Layer Semantic Network. *Applied Sciences*, 11(3).

[145] Zhu, G. and Iglesias, C. A. (2017a). Computing Semantic Similarity of Concepts in Knowledge Graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85.

[146] Zhu, G. and Iglesias, C. A. (2017b). Sematch: Semantic Similarity Framework for Knowledge Graphs. *Knowledge-based Systems*, 130:30–32.

Abstract

There is a widespread of information published every day on the internet. there are a lot of books from different cultures that are now digitized and kept in digital media in the form of natural language. this means that it is key to have solutions for extracting important or useful information from the massive amount of data. Computing sentence similarity and the performance of such systems is a delicate matter of huge potential. Semantic search, extractive summarization, sentiment analysis, document classification, plagiarism detection can all be considered special cases of similarity. Sentence similarity assesses the similarity of phrases in a sentence pair. The topic of this research is sentence similarity assessment, which is composed of short texts of less than two lines. In the case of sentence pair similarity, two sentences are sent to the system and the system must produce an output of range zero to one (zero means the pair share no similarity, and one means they are semantically equivalent). The purpose of this research is to design a system that can have a better correlation with those of human judges and to study the key elements of state-of-the-art systems.

In the proposed wavelet transform method, first, we fine-tune the sentence encoder on the corresponding training data. Next, we decompose ”sentence description” using wavelet transform, next, the cosine similarities of the corresponding channels of descriptors are used to compute sentence similarity. The results suggest significant improvement over the Roberta-base method and all other base encoders. Most superior methods use a large model encodes however, in comparison our approach uses a base model which means less training and test time.

Key Words: Text similarity, Word co-occurrence, perplexity, word pair similarity, regression.

Shahrood University of Technology

Faculty of Computer Engineering and Information Technology

Ph.D. Thesis in Artificial Intelligence

Improving Sentence Similarity Measures Using Statistical Approaches and WordNet-based Metrics

By: Reza Javadzadeh

Supervisor:
Morteza Zahedi

Advisor:
Marziea Rahimi

June 2021