Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 11, Number 1, Spring 2022, pp. 18-31
Received: 08/18/2021, Revised: 01/07/2022, Accepted: 04/04/2022

Semantic Textual Similarity of Persian-English sentences using deep learning

Mohammad Abdous[1] and Behrouz Minaei Bidgoli[2*]

1- Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.

2*- Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.

[1]Mohammadabdous@comp.iust.ac.ir and [2*]B_Minaei@iust.ac.ir

Corresponding author's address: Behrouz Minaei Bidgoli, Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.

Abstract- Semantic Textual similarity is one of the subtasks of natural language processing that has attracted extensive rese arch in recent years. Measuring semantic similarity between words, sentences, paragraphs, and documents plays an important role in natural language processing and computational linguistics. Semantic similarity of texts is used in question-answering systems, fraud detection, machine translation, information retrieval and etc. Semantic similarity means calculating the degree of similarity between two textual documents, paragraphs or sentences, which are presented in both monolingual and cross lingual forms. In this article, by using the parallel corpus, for the first time, the cross lingual model of semantic similarity for Persian-English sentences is presented, and then we test and compare our model with the Multilingual BERT model. The results show that by using parallel corpuses, the quality of sentence embedding in two different languages can be improved. Pearson correlation criterion based on cosine similarity between sentence's vector of multilingual Bert has increased from 65% to 73.77% by the proposed method. The proposed method was also tested on the Arabic-English language pair, and the results show that the proposed method is superior to the multilingual Bert.

Keywords- Natural language processing, semantic similarity, cross lingual, deep learning

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۱، شماره ۱، بهار ۱۴۰۱ ، صفحه ۱۸-۳۱
دریافت: ۱۴۰۰/۰۵/۲۷، بازنگری: ۱۴۰۰/۱۰/۱۷، پذیرش: ۱۴۰۱/۰۱/۱۵

شباهت‌یابی بین زبانی جملات فارسی-انگلیسی با استفاده از یادگیری عمیق

محمد عبدوس[۱] و بهروز مینایی بیدگلی*[۲]

۱- دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران.

*۲- دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران.

[1]mohammadabdous@comp.iust.ac.ir, [2*]b_minaei@iust.ac.ir

* نشانی نویسنده مسئول: بهروز مینایی بیدگلی، تهران، دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر.

چکیده-شباهت‌یابی معنایی متون یکی از زیرشاخه‌های پردازش زبان طبیعی محسوب می‌شود که در چند سال اخیر تحقیقات گسترده‌ای را به خود معطوف کرده است. سنجش تشابه معنایی بین کلمات یا اصطلاحات، جملات، پاراگراف و اسناد، نقش مهمی در پردازش زبان طبیعی و زبان‌شناسی رایانشی ایفا می‌کند. شباهت‌یابی معنایی متون در سامانه‌های پرسش و پاسخ، کشف تقلب، ترجمه ماشینی، بازیابی اطلاعات و نظیر آن کاربرد دارد. منظور از شباهت‌یابی معنایی، محاسبه میزان شباهت معنایی بین دو سند متنی، پاراگراف یا جمله می‌باشد که به دو صورت تک‌زبانه و چندزبانه مطرح است. در این مقاله با استفاده از پیکره موازی میزان، برای اولین بار مدل بین زبانی شباهت معنایی جملات فارسی-انگلیسی را ارائه داده و در ادامه مدل خود را با مدل برت چندزبانه مورد آزمون و مقایسه قرار دادیم. نتایج حاکی از آن است که با استفاده از پیکره‌های موازی می‌توان کیفیت تعبیه جملات را در دو زبان مختلف بهبود بخشید. در روش پیشنهادی، معیار همبستگی پیرسون بر اساس شباهت کسینوسی بین بردارهای معنایی حاصل از برت چندزبانه از ۶۵ درصد به ۷۳.۷۷ درصد افزایش یافته است. روش پیشنهادی بر جفت زبان عربی-انگلیسی نیز مورد آزمون قرار گرفت که نتایج حاصله بیانگر برتری روش پیشنهادی نسبت به برت چند زبانه است.

واژه‌های کلیدی: پردازش زبان طبیعی، شباهت معنایی، بین زبانی، یادگیری عمیق

۱- مقدمه

امروزه اطلاعات متنی در بسترهای مختلف نظیر اینترنت، مجلات و مانند آن به‌صورت روزافزون در حال گسترش است و حجم عظیمی از داده‌های متنی هر لحظه اضافه می‌شود. علاوه بر این، کتاب-خانه‌های دیجیتالی نیز با روند رو به رشدی در حال توسعه است. حجم بسیار زیادی از این داده‌ها در ساختار زبان طبیعی قالب‌بندی شده‌اند. همین امر باعث می‌شود تا تکنیک‌های پردازش زبان طبیعی برای استفاده از این حجم داده بسیار مهم باشند. پردازش زبان طبیعی[۱] یکی از پرکاربردترین حوزه‌های هوش مصنوعی[۲] است و الگوریتم‌های زیادی توسط پژوهشگران جهت حل مسائل آن ارائه شده است. با پیشرفت روزافزون توان سخت‌افزاری و پردازشی برای آموزش مدل‌های جدید و عمیق، توجه جامعه علمی بیش‌ازپیش بر استفاده و بهبود ابزارهای موجود جهت حل مسائل مختلف معطوف شده است.

محاسبه‌ی میزان شباهت معنایی بین بخش‌های متنی (کلمات، جملات، پاراگراف‌ها یا حتی اسناد) یک زمینه تحقیقاتی بسیار مهم در پردازش زبان طبیعی است. شباهت‌یابی معنایی بین جملات در بسیاری از کاربردهای زبان طبیعی مانند جستجوی معنایی[۳][۱]،

خلاصه سازی[۴][۲]، سامانه‌های پرسش و پاسخ[۵][۳]، رده‌بندی اسناد[۶][۴]، تحلیل احساسات[۷][۵] و سرقت ادبی[۸][۶] مورد استفاده قرار می‌گیرد. تشخیص میزان شباهت معنایی با هدف درک و تولید زبان انسانی، پژوهشی جذاب در علوم کامپیوتر، هوش مصنوعی و زبان‌شناسی محاسباتی است. میزان تشابه بین جملات با درجه احتمال تشابه بین آن‌ها نشان داده می‌شود. شباهت‌یابی معنایی از سال ۲۰۱۲ در کنفرانس ارزش‌یابی معنایی مورد توجه قرار گرفت و پس از آن با توجه به اهمیت این کار، هر ساله کنفرانسی با نام شباهت‌یابی معنایی برگزار می‌شود. شباهت‌یابی معنایی در سال‌های ۲۰۰۶ تا ۲۰۱۲ به‌عنوان یک مسأله رده‌بندی دو کلاسه مورد توجه قرار گرفته بود (تشخیص اینکه دو جمله شباهت معنایی دارند یا خیر)، اما از سال ۲۰۱۲ به بعد محاسبه میزان شباهت که با عدد بیان می‌شود، در دستور کار قرار گرفت[۷].

امروزه محققان با استفاده از مدل‌های یادگیری عمیق، جانمایی‌های مختلفی از کلمات و جملات را ارائه داده‌اند. در بسیاری از روش‌های مطرح یک کلمه یا جمله به یک بردار عددی نگاشت پیدا می‌کند. عملیات نگاشت یک کلمه یا جمله به یک بردار عددی را جانمایی می‌گویند. در روش‌های اولیه نگاشت، به معنای کلمه یا جمله توجهی نمی‌شد اما با مرور زمان پژوهشگران دریافتند که با استفاده از مدل‌های یادگیری عمیق می‌توان بردارهای مختلف حاوی معنا را تولید کرد. معنادار بودن بردارهای تولید شده می‌تواند در بسیاری از کارها مورد استفاده قرار گیرد؛ چراکه امروزه متخصصان پردازش زبان طبیعی به سراغ مباحثی پیرامون درک و فهم زبان طبیعی رفته‌اند و این مهم جز با جانمایی معنایی از جملات یا کلمات میسر نیست. جانمایی‌های مختلف که جهت تولید بردار معنایی جملات ایجاد می‌شوند، غناهای معنایی متفاوتی دارند. برخی از این جانمایی‌ها تنها استخراج بردار کلمه را مدنظر قرار داده و برخی دیگر با استفاده از روش‌هایی نظیر یادگیری عمیق، بردار معنایی جملات را تولید کرده‌اند.

میزان شباهت بین جملات با تبدیل جمله به بردارهای معنایی میسر است. با بهره‌گیری از روش‌های جانمایی کلمات[۹] یا جملات، آن‌ها در یک فضای برداری نگاشت شده و میزان شباهت معنایی بین جملات با استفاده از آن محاسبه می‌شود. معیارهای مختلفی جهت محاسبه میزان شباهت وجود دارد که از معروف‌ترین آن‌ها می‌توان به فاصله اقلیدسی[۱۰]، فاصله منهتن[۱۱] و شباهت کسینوسی[۱۲] بین دو بردار جملات اشاره کرد. درجه‌بندی میزان شباهت معمولا بین صفر تا یک است اما در مجموعه داده‌های محک شباهت‌یابی معنایی[۱۳]، میزان شباهت از ۰ تا ۵ درجه‌بندی شده است.

جملات می‌توانند از نظر لغوی یا معنایی مشابه باشد. تشابه لغوی(واژگانی) به معنای شباهت مبتنی بر رشته و شباهت معنایی بیانگر وجود معنا و مفهوم مشابه بین جملات است؛ حتی اگر کلمات مورد استفاده در جملات متفاوت باشند. از این تعریف، رویکردهای ارائه شده می‌تواند به شباهت مبتنی بر رشته و شباهت معنایی طبقه‌بندی شوند.

رویکرد شباهت مبتنی بر رشته یا مبتنی بر واژگان، جمله را دنباله ای از نویسه‌ها[۱۴] در نظر می‌گیرد. محاسبه شباهت در این روش به اندازه‌گیری تشابه بین توالی نویسه‌ها بستگی دارد. روش‌های بسیاری مانند فاصله لونشتاین[۱۵]، همینگ[۱۶] و ضریب شباهت جاکارد[۱۷] جهت محاسبه میزان شباهت بین دو رشته وجود دارد که فقط به ظاهر کلمه توجه دارند و معنای آن را در نظر نمی-گیرند[۸]، [۹].

امروزه اکثر روش‌های شباهت معنایی مبتنی بر شبکه‌های عصبی عمیق[۱۸] هستند [۱۰] و تمرکز اصلی مقاله به آن اختصاص دارد. در این رویکرد از روش‌های معنایی جهت محاسبه میزان شباهت بین جملات استفاده می‌کند. این رویکردها با استفاده از پیکره‌های بزرگ، کلمات را در فضای برداری نگاشت می‌کنند و میزان شباهت را بر اساس نزدیکی در فضای برداری می‌سنجند که معمولا از روش‌های یادگیری عمیق در این رویکرد استفاده می‌شود[۱۱]. همچنین محققان از پایگاه‌های دانش که گراف معنایی کلمات در آن ایجاد شده، جهت محاسبه شباهت معنایی نیز استفاده می‌کنند[۱۲]. استفاده از پایگاه‌های دانش نیز در زمره روش‌های مربوط به شباهت معنایی قرار می‌گیرد.

جملات در زبان‌های مختلف نیز می‌تواند از لحاظ معنایی مشابه هم باشند. به‌عنوان نمونه جمله "He wants to play football" و جمله "احمد در ورزشگاه تختی تمرین می‌کند" می‌تواند از لحاظ معنایی شبیه به هم باشد. این دو جمله در دو زبان با ساختار متفاوت قرار دارند. در بسیاری از کارهای انجام شده جهت محاسبه میزان شباهت بین دو زبان مختلف از ماشین ترجمه استفاده می‌کنند و جمله زبان مبدا را به زبان مقصد ترجمه کرده و سپس از مدل‌های مربوط به زبان مقصد(در مثال فوق انگلیسی) جهت محاسبه میزان شباهت معنایی استفاده می‌کنند؛ اما ضعف اصلی چنین سامانه‌هایی انتشار خطای ماشین ترجمه است و ممکن است

ترجمه زبان مبدا به زبان مقصد به خوبی صورت نپذیرد. در جدول ۱ نمونه جمله فارسی و جمله انگلیسی را به همراه میزان شباهت معنایی آن که در بازه ۰ تا ۵ قرار دارد، مشاهده می‌کنید. امتیاز ۵ بیانگر بیشترین میزان شباهت (برابری کامل معنایی بین دو متن) و ۰ بیانگر کمترین میزان شباهت (عدم ارتباط معنایی بین دو متن) است.

جدول ۱- نمونه جمله فارسی - انگلیسی به همراه میزان شباهت معنایی

جمله فارسی	جمله انگلیسی	میزان شباهت
اعتراضات در پایتخت اوکراین ادامه دارد.	Protests continue in tense Ukraine capital.	۵
ناسا قرار است از صبح دوشنبه تاسیسات تلسکوپ مادون قرمز فضایی را راه‌اندازی کند.	The Space Infrared Telescope Facility's mission is to search for the beginnings of the universe.	۲.۵

در این مقاله با استفاده از پیکره‌ی موازی فارسی-انگلیسی میزان، فضای مدل‌های مبتنی بر مبدل‌ها نظیر برت چندزبانه را تغییر داده و امکان استفاده از آن مدل را در حالت بین زبانی فراهم نمودیم.

مهم‌ترین هدف این مقاله آن است که باید دو جمله مشابه از لحاظ معنایی در دو زبان مختلف، بردار جملات نزدیک به هم را در فضای برداری داشته باشند تا میزان شباهت معنایی آن‌ها را بتوان با استفاد از معیارهای مختلف سنجید. چالش اصلی در مدل‌های چندزبانه عدم وجود فضای مشترک برداری بین زبان‌هایی با ساختار مختلف است. به همین علت استفاده از این مدل‌ها جهت محاسبه شباهت معنا بین دو جمله در دو زبان مختلف عملکرد خوبی را ارائه نمی‌دهد. هرچه بردارهای جملات از لحاظ مفهومی شبیه‌تر باشند، باید در فضای برداری به هم نزدیکتر بوده و بالطبع سامانه شباهت‌یابی معنایی نیز از دقت بالاتری برخوردار خواهد بود. بنابراین یکی از مهم‌ترین نکات مورد توجه در بحث شباهت‌یابی معنایی جملات، نحوه جانمایی بردارهای جملات است. در این مقاله با توجه به برخی از آزمایش‌های صورت پذیرفته نشان می‌دهیم که شبکه‌های عصبی عمیق می‌توانند چالش فضای مشترک برداری بین دو زبان با دو ساختار متفاوت را حل کنند و می‌توان از مدل ایجادشده جهت شباهت‌یابی معنایی بین زبانی استفاده نمود. در شباهت‌یابی معنایی چندزبانه جملاتی که از لحاظ معنایی شبیه به هم هستند باید بردارهای نزدیک به هم داشته باشند که در مدل‌های مطرح چندزبانه مانند برت‌چندزبانه[۱۹] برای زبان‌های با ساختار متفاوت(مانند انگلیسی و فارسی) این‌گونه نیست[۱۳]. استفاده از دادگان موازی[۲۰]، ما را قادر می‌سازد تا جانمایی‌های با کیفیتی را در زبان‌های مختلف آموزش دهیم و نحوه جانمایی‌های مدل‌های چندزبانه مانند برت‌چندزبانه را تغییر دهیم. شباهت‌یابی معنایی بین زبانی متون کاری جدید و ارزشمند است که متاسفانه تاکنون در زبان فارسی به آن پرداخته نشده است. همچنین وجود مدل‌های جدید زبانی که بتواند جانمایی‌های نزدیک به هم را در زبان‌های مختلف داشته باشد از دیگر نوآوری‌های طرح پیشنهادی است.

در مجموع، نوآوری‌های این مقاله عبارتند از:

۱. تغییر فضای برداری مدل برت چندزبانه با استفاده از پیکره‌های موازی
۲. تولید مدل شباهت‌یابی معنایی بین زبانی فارسی انگلیسی برای اولین بار با استفاده از یادگیری عمیق

در ادامه این مقاله در بخش دوم کارهای مرتبط شباهت یابی معنایی مورد بررسی قرار میگیرد. در بخش سوم روش پیشنهادی شرح داده می‌شود و سپس نتایج و آزمایش‌های صورت پذیرفته بیان می‌گردد.

۲- کارهای مرتبط

محاسبه شباهت بین متون کوتاه اولین بار در سال ۲۰۰۶ گزارش شد[۱۴]، [۱۵]. پس از آن از سال ۲۰۱۲ در کارگاه بین‌المللی ارزیابی معنایی، وظیفه شباهت معنایی تنها به وجود یا عدم وجود شباهت معنایی محدود نشد بلکه برای هر جفت متن (جمله) محاسبه میزان شباهت که عددی بین ۰ تا ۵ است مورد تمرکز قرار گرفت[۱۶]–[۱۸]. این کارگاه به جنبه‌های مهم پردازش زبان طبیعی و هوش مصنوعی می‌پردازد که شباهت‌یابی معنایی یکی از آن‌هاست. در شباهت‌یابی معنایی میزان شباهت بین دو جمله با عدد مشخص می‌شود که معمولاً عددی بین ۰ تا ۵ است. ایده‌های اولیه برای شناسایی شباهت معنایی بین دو جمله مبتنی بر هم‌ترازی معنایی بین کلمات دو جمله و درنهایت جمع جبری بین شباهت‌های بین کلمات بود [۱۹] اما امروزه بیشتر پژوهش‌ها در زمینه جانمایی معنایی یک جمله با استفاده از یادگیری عمیق متمرکز است. جملات با استفاده از این‌گونه روش‌ها به بردارهای

عددی با ابعاد مختلف تبدیل می‌شوند که حامل معانی کلمات در فضای برداری هستند و کلماتی که در آن فضا به هم نزدیک‌ترند، از لحاظ معنایی نیز شباهت دارند. تولید بردار معنایی کلمات با استفاده از پیکره‌های بزرگ متنی صورت می‌پذیرد. در زبان انگلیسی به علت وسعت کشورهای انگلیسی‌زبان و همچنین فراگیر بودن آن در تمام دنیا پژوهش‌های بیشتری در این زمینه انجام شده اما در زبان‌هایی که از منابع و پیکره‌های محدودتری برخوردارند مانند زبان فارسی پژوهشی در این زمینه صورت نپذیرفته است. با توجه به تعداد زیاد پژوهش‌ها در زمینه شباهت‌یابی معنایی اعم از تک‌زبانه و چندزبانه، در این بخش تنها به بیان کارهای مرتبط چندزبانه شباهت‌یابی معنایی پرداخته می‌شود.

در شباهت‌یابی معنایی چندزبانه، هدف محاسبه میزان شباهت بین جملات دو زبان مختلف است. در زبان‌هایی که از منابع داده‌ای مناسبی برخوردارند، مشکلی در محاسبه میزان شباهت معنایی وجود ندارد اما در برخی از زبان‌ها که منبع مناسبی وجود ندارد محاسبه میزان شباهت با چالش جدی مواجه است. یکی از راه حل‌های موجود برای حل این چالش استفاده از رویکردهای مبتنی بر ترجمه ماشینی برای تبدیل جملات از زبان کم‌منبع مانند فارسی به زبان با منبع زیاد مانند انگلیسی است. مشکلی اصلی چنین رویکردهایی وجود خطا در ترجمه ماشینی است و به شدت به کیفیت ترجمه وابسته است[۲۰]. در ادامه به معرفی پژوهش‌های برتر شباهت‌یابی بین زبانی پرداخته شده است.

تانگ[۲۱] و همکاران [۲۱] در پژوهش خود مدلی را برای زبان‌های کم‌منبع مانند اسپانیایی، عربی، اندونزیایی و تایلندی ارائه داده‌اند. آن‌ها با استفاده از چهارچوب مدل شباهت‌یابی معنایی تک‌زبانه، اقدام به گسترش آن در حالت چندزبانه نموده و نشان دادند که با استفاده از یک رمزگذار مشترک چندزبانه، هر جمله می‌تواند جانمایی‌های مختلفی را با توجه به زبان هدف از خود نشان دهد.

بریچین[۲۲] [۲۲] ایده‌ای را مطرح کردند که در آن فضاهای معنایی چندزبانه با استفاده از فرهنگ لغت‌های دوزبانه در یک فضای مشترک قرار می‌گیرند. آن‌ها از روش‌های بدون نظارت برای شباهت جملات فقط بر اساس فضاهای معنایی سامانه‌ای را ایجاد کرده و نشان دادند که می‌توان فضاهای مشترک معنایی را با وزن‌دهی به کلمات بهبود بخشید. نتایج آن‌ها بیانگر معیار همبستگی پیرسون ۶۱.۸ درصد در جفت جملات عربی – انگلیسی است.

در کنفرانس ارزشیابی معنایی سال ۲۰۱۷ تمرکز اصلی بر شباهت‌یابی معنایی بین زبانی و چند زبانی بود[۲۳]. در این کنفرانس ۱۷ شرکت‌کننده در قالب ۳۱ تیم به رقابت پرداختند و در این کنفرانس پیکره محک شباهت‌یابی معنایی (STSBenchMark) ارائه گردید. همچنین برخی از پیکره‌های بین زبانی عربی-انگلیسی، اسپانیایی-انگلیسی و ترکی-انگلیسی نیز معرفی شد و توسط شرکت‌کنندگان مورد ارزیابی قرار گرفت. در ادامه به برخی از بهترین کارهای این کنفرانس اشاره می‌شود.

مدل ECNU که توسط تیان[۲۳] و همکاران ارائه گردیده است [۲۴] رتبه اول را در بین شرکت‌کنندگان کنفرانس ارزشیابی معنایی ۲۰۱۷ به‌دست آورده است. این مقاله با اینکه رتبه اول را به خود اختصاص داده اما مدل بین زبانی را ارائه نداده و از ماشین ترجمه گوگل استفاده کرده و جملات زبان غیر انگلیسی را به انگلیسی ترجمه کرده و سپس یک مدل تک‌زبانه ایجاد شده که با استفاده از سه ماژول مختلف اقدام به شباهت‌یابی معنایی جملات انگلیسی-انگلیسی نموده‌اند. ماژول سنتی NLP که برای استخراج دو نوع از ویژگی‌های پردازش زبان طبیعی به کار می‌رود. برخی از ویژگی‌ها مانند همپوشانی ان-گرم‌ها که از جفت جمله ورودی به‌دست می‌آید و برخی از ویژگی‌ها مانند کیف کلمات که وابسته به خود جمله است. پس از استخراج این دو نوع ویژگی مدل‌هایی مبتنی بر الگوریتم‌های RF^{24},GB^{25} و XGB^{26} ایجاد شده است. ماژول یادگیری عمیق با استفاده از بردارهای ورودی کلمات اقدام به ایجاد مدل نموده است. بردارهای کلمات از ۴ منبع مختلف word2vec[11], GloVe 100d,GloVe 300d [25] و[۲۶] paragramبه‌دست آمده‌اند. سپس با استفاده از بردارهای کلمات، جانمایی‌های مختلفی از جملات استخراج نمودند. بردار جملات با استفاده از میانگین بردار کلمات، میانگین بردار کلمات پیش‌بینی شده[۲۷]، شبکه DAN^{28} و شبکه LSTM ایجاد شده‌اند. در ادامه از اتصال تفریق بردار دو جمله و ضرب آن‌ها جهت به‌دست آوردن بردار جفت جمله استفاده می‌گردد. در نهایت، از یک شبکه عصبی کاملاً متصل استفاده شده و احتمال شباهت بر اساس تابع softmax به‌دست می‌آید. بنابراین در این مرحله ۴ عدد میزان شباهت مبتنی بر یادگیری عمیق تولید می‌گردد.

در پایان با استفاده از میانگین امتیازهای حاصل از مدل‌های سنتی پردازش زبان طبیعی و امتیازهای به‌دست آمده در ماژول یادگیری عمیق، میزان شباهت معنایی جفت جمله به‌دست می‌آید. بهترین نتیجه مربوط به ترکیب روش سنتی با روش یادگیری عمیق است

که از میانگین هفت امتیاز به‌دست آمده است. این ۷ امتیاز شامل ۳ امتیاز روش‌های سنتی و ۴ امتیاز روش‌های یادگیری عمیق است.

وو[۲۹] و همکاران [۲۷] سه سامانه را جهت شباهت‌یابی معنایی ارائه داده‌اند و مقام دوم را در کنفرانس شباهت‌یابی معنایی ۲۰۱۷ کسب کرده‌اند. یک سامانه بدون نظارت و دو سامانه دیگر با ناظر هستند که هر سه سامانه به فضای اطلاعات معنایی[۳۰] وابسته هستند. این فضا براساس طبقه‌بندی معنایی سلسله مراتبی موجود در وردنت[۳۱] ایجاد شده است. سامانه بدون نظارت با استفاده از محتوای اطلاعاتی بدون همپوشانی در فضای اطلاعات معنایی استفاده می‌کند و دو سامانه دیگر با استفاده از هم‌ترازی جملات[۳۲] و جانمایی کلمات اقدام به شباهت‌یابی معنایی جملات می‌کنند. ترکیب ویژگی محتوای اطلاعاتی با شباهت کسینوسی بردار جملات بهترین عملکرد را دارد. این تیم طبق ارزیابی به‌دست آمده با ضریب همبستگی پیرسون در بین ۳۱ تیم شرکت کننده رتبه دوم را از آن خود کرده و بهترین عملکرد را در مجموعه داده محک شباهت‌یابی معنایی عربی-عربی کسب کرده است.

شاو[۳۳] [۲۸] با استفاده از یادگیری عمیق اقدام به تولید مدلی با نام HCTI جهت شباهت‌یابی معنایی نموده است. این سامانه از ۵ قسمت به شرح زیر تشکیل شده است. ۱) بردار کلمات جمله که از مجموعه بردارهای گلاو گرفته شده با ویژگی‌هایی نظیر برچسب ادات سخن[۳۴] کلمه ترکیب شده است. ۲) هر کدام از بردارهای ایجاد شده در مرحله قبل به یک شبکه عصبی پیچشی داده شده و خروجی آن به‌عنوان بردار جدید کلمه معرفی می‌گردد.۳) جهت تولید بردار معنایی جمله، در هر بعد بردار کلمات عملیات max pooling انجام شده و بردار نهایی جمله به‌دست می‌آید.۴) از اتصال تفریق و ضرب دو بردار جمله، بردار معنایی جدیدی تولید می‌شود که بیانگر اختلاف معنایی دو جمله است۵) بردار اختلاف معنایی تولید شده در مرحله قبل به یک شبکه عصبی عمیق تماما متصل داده شده و خروجی آن به یک لایه softmax داده شده تا میزان احتمال مربوط به هر امتیاز (عدد ۰ تا ۵) محاسبه گردد. این سامانه در کنفرانس ارزشیابی ۲۰۱۷ رتبه سوم را به‌دست آورده است.

کارهای دیگری که در زمینه شباهت معنایی بین‌زبانی انجام شده مبتنی بر جانمایی‌های بین زبانی است که از مهم‌ترین آن‌ها می‌توان به [۲۹]–[۳۳] اشاره کرد. چیدامبارام[۳۵] و همکاران [۳۴] فضای برداری بین زبانی را با استفاده از مدل مبتنی بر رمزگذار دوتایی تولید کرده‌اند. هدف آموزش مدلی است که بیشینه شباهت را بین جفت جملات موجود در پیکره موازی تولید کند. جانمایی‌های به‌دست آمده با استفاده از پیکره‌های تک‌زبانه و آموزش همزمان چند وظیفه، بهبود پیدا می‌کنند. آن‌ها از فضای برداری مشترک به‌دست آمده در بسیاری از کارها استفاده نموده و در مقایسه با سایر کارها از برتری نسبی برخوردار است. هسته اصلی روش پیشنهادی آن‌ها مدل‌سازی وظایفی است که مبتنی بر رتبه‌بندی جفت جملات با استفاده از رمزگذارهای دوگانه است. جانمایی‌های بین‌زبانی با در نظرگرفتن یک کار ماشین ترجمه به‌دست می‌آید. در معماری رمزگذار اشتراکی سه زیرشبکه مبدل وجود دارد که هریک دارای زیرلایه‌های روبه‌جلو و توجه چندراسی است. خروجی مبدل یک توالی با طول متغیر است که با میانگین‌گیری آن‌ها، جانمایی جملات به‌دست می‌آید. جانمایی‌های ایجادشده سپس در مجموعه‌های مختلفی از لایه‌های روبه‌جلو که برای هر کار استفاده می‌شوند، قرار می‌گیرد.

کنو[۳۶] و همکاران [۳۵] مجموعه داده بین زبانی با نام XNLI را تولید کرده‌اند. از آنجا که جمع‌آوری داده‌ها به تمامی زبان‌ها فرآیندی هزینه‌بر است علاقه‌مندی به درک بین زبانی[۳۷] و انتقال به زبان‌های با منبع کم افزایش یافته است. در این مقاله، یک مجموعه ارزیابی برای درک بین‌زبانی ساخته شده و مجموعه‌های آزمایشی به ۱۵ زبان گسترش یافته است از جمله زبان‌های کم‌منبع مانند سوحیلی[۳۸] و اردو. در پیکره تولید شده جملات در زبان‌های مختلف دارای دو فرضیه premise و hypothesis هستند. آن‌ها برای اثبات ارزشمندبودن مجموعه داده، آن را بر روی چند کار مانند ترجمه ماشینی، کیف کلمات چندزبانه و رمزگذار LSTM مورد آزمایش قرار دادند.

کنو و همکاران [۳۶] مدل بین‌زبانی با نام XLM را ارائه داده‌اند. آن‌ها دو روش را برای یادگیری مدل‌های بین‌زبانی به کار بستند. روش اول بدون نظارت که تنها به داده‌های تک‌زبانه متکی است و روش دوم از پیکره موازی با هدف مدل‌سازی زبان مقصد به صورت باناظر استفاده می‌کند. روش پیشنهادی آن‌ها روی وظایف ترجمه ماشینی باناظر و بدون‌ناظر و رده‌بندی بین‌زبانی(XNLI) بهترین عملکرد را از آن خود کرده است. هدف مدل زبانی نقاب‌دار[۳۹] مشابه هدفی است که توسط داولین در مقاله برت [۳۷] معرفی شده است با این تفاوت که در این مدل با جریان مداوم جفت جمله روبرو هستیم. در مدل زبان ترجمه همانند مدل زبانی نقابدار جفت جملات موازی به ماشین داده می‌شود. برای پیش‌بینی یک

کلمه انگلیسی، این مدل می‌تواند هم به جمله انگلیسی و هم به ترجمه فرانسوی آن توجه داشته باشد، و رویکرد آن این است که جانمایی‌های انگلیسی و فرانسوی را هم‌تراز کند.XLM به منظور یادگیری روابط بین کلمات در زبان‌های مختلف، از روش رمزگذاری جفت بایت[۴۰] و سازوکار آموزش دو زبان با برت استفاده می‌کند. رمزگذاری جفت بایت یک روش فشرده‌سازی داده است که به طور مداوم پرتکرارترین جفت نویسه (در اصل بایت) را در یک مجموعه داده خاص با یک نماد بدون رخداد در متن جایگزین می‌کند. در هر تکرار، الگوریتم پرتکرارترین جفت نویسه را پیدا می‌کند و آن‌ها را ادغام می‌کند تا نماد جدیدی ایجاد شود. در مدل XLM، به جای استفاده از کلمه یا کاراکترها به‌عنوان ورودی مدل، از رمزگذاری جفت بایت استفاده می‌کند که ورودی را به متداول‌ترین زیرکلمه در تمامی زبان‌ها تقسیم می‌کند و با این کار واژگان مشترک بین زبانی افزایش پیدا می‌کند. مدل XLM معماری برت را به دو روش ارتقا می‌دهد: در مدل برت هر نمونه آموزشی از یک زبان تشکیل شده در حالی‌که در مدل XLM هر نمونه آموزشی متشکل از دو زبان است. همانند پیش‌بینی کلمات نقاب‌دار در مدل برت در این مدل از بافتار جمله مبدا برای پیش‌بینی کلمات نقاب جمله مقصد استفاده می‌کند. همچنین این مدل شناسه زبان و ترتیب کلمات در هر زبان، به‌عنوان رمزگذار موقعیت، به طور جداگانه دریافت می‌کند. فراداده جدید به مدل کمک می‌کند تا رابطه بین کلمات مرتبط در زبان‌های مختلف را بیاموزد.

سور[۴۱] و ارکان[۴۲] [۳۸] جهت ارزیابی روش‌های شباهت معنایی بین زبانی از تعاریف وردنت در ۷ زبان مختلف استفاده نمودند. مجموعه ترادف‌های[۴۳] وردنت در ۷ زبان مختلف را هم‌تراز نموده و مقایسه‌ای بین روش‌های باناظر و بی‌ناظر ارائه نموده‌اند. همچنین به این نتیجه رسیده‌اند که با استفاده از شباهت‌یابی معنایی بین زبانی می‌توان به صورت خودکار وردنت را ایجاد کرد.

وانگ و همکاران [۳۹] با استفاده از یادگیری متضاد[۴۴] روشی را پیشنهاد نمودند که در آن روش هی و همکاران [۴۰] با نام موکو برای ایجاد فضای تعبیه مشترک و هم‌ترازی جملات بهبود پیدا کرده است. آن‌ها از برت استفاده کرده و جانمایی‌هایی برای زبان چینی و انگلیسی به دست آورده‌اند.

بریاکو[۴۵] و کارپوت[۴۶] [۴۱] معتقدند تشخیص تفاوت‌ها در محتواهای زبان‌های مختلف جهت وظایف پردازش زبان طبیعی از اهمیت ویژه‌ای برخوردار است. اما چالش اصلی آن هزینه‌بر بودن حاشیه‌نویسی است. آن‌ها با پژوهش خود کار پیش‌بینی و حاشیه‌نویسی تفاوت‌های معنایی را بهبود بخشیده‌اند. آن‌ها یک استراتژی آموزشی برای مدل برت چندزبانه ارائه داده و رده‌بند دودویی (کلاس هم‌ارزی[۴۷] و کلاس واگرایی[۴۸]) بر اساس ورودی جمله انگلیسی و جمله فرانسوی آموزش داده‌اند. لایه رده‌بند بر بالاترین لایه برت چند زبانه قرار گرفته است.

دوتا[۴۹] در پژوهشی [۴۲] با تاکید بر اینکه شباهت‌یابی بین زبانی متون معیار مهمی برای مقایسه بین اسناد در دو زبان است، فضای برداری بین زبانی جدیدی را با استفاده از فاصله word Mover در تعبیه هم‌تراز شده جملات به دست آورده‌اند. نتایج وی بیانگر آن است که استفاده از این فاصله می‌تواند روشی بهینه و بی‌ناظر جهت محاسبه میزان شباهت معنایی بین دو زبان باشد.

در ادامه توضیحاتی پیرامون عملکرد برت چندزبانه در مورد زبان‌های با منابع کم ارائه خواهیم داد. با توجه به اهمیت برت چندزبانه و عملکرد مناسب آن، پژوهشگران اقدام به تولید مدل چندزبانه با نام برت چندزبانه نمودند. این مدل با استفاده از متن خام ویکی‌پدیا در ۱۰۴ زبان مختلف تولید شده است. این مدل در فرآیندی کاملا بدون نظارت آموزش می‌بیند و از هیچ‌گونه داده موازی نیز در فرآیند تولید آن استفاده نشده است اما با این حال در برخی از زبان‌ها تعمیم بین‌زبانی دارد[۴۳]. پیرز[۵۰] و همکاران [۱۳] در مقاله‌ای به بررسی کیفیت برت چندزبانه برای کارهای بین‌زبانی پرداختند. آن‌ها آزمایش‌های مختلفی بر پیکره‌های مختلف با استفاده از مدل برت چندزبانه انجام داده و به نتایج خوبی دست پیدا کرده‌اند. در برخی از آزمایش‌ها که بر روی دو زبان مختلف انجام شده، جانمایی‌های بین‌زبانی برای جفت جملات برخی زبان‌ها مانند انگلیسی و ژاپنی از دقت کمی برخوردار است و این کمی دقت به علت ساختارهای متفاوت دو زبان است. در این مقاله زبان‌ها را به دو دسته تقسیم کردند. برخی از زبان‌ها از لحاظ قواعد گرامری، گرامر متفاوتی نسبت به زبان انگلیسی دارند. زبان‌ها را می‌توان از لحاظ گرامری به دو دسته SVO^{51} , SOV^{52} تقسیم‌بندی کرد. زبان‌هایی مانند انگلیسی SVO هستند یعنی ترتیب قرارگیری بخش‌های جمله (در یک جمله ساده یا خبری) به صورت فاعل، فعل بعد از آن و در نهایت مفعول است. زبان‌هایی مانند فارسی SOV هستند یعنی در فارسی بر خلاف انگلیسی ترکیب قرارگیری اول فاعل، بعد از آن مفعول و در نهایت فعل (معمولا در پایان جمله) است. در شکل ۱ درصد دقت برچسب زنی ادات سخن برت چندزبانه در زبان‌های با ساختار متفاوت

نمایش داده شده است. منظور از عبارت NA[53] در شکل ۱ ابتدا اسم و سپس صفت (مانند "درخت زیبا" در زبان فارسی) و منظور از AN[54] ابتدا صفت و سپس اسم(مانند "beautiful tree" در زبان انگلیسی) است.

	SVO	SOV
SVO	**81.55**	66.52
SOV	63.98	**64.22**

(a) Subj./verb/obj. order.

	AN	NA
AN	**73.29**	70.94
NA	75.10	**79.64**

(b) Adjective/noun order.

شکل ۱-دقت برچسب‌زنی اجزای سخن در برت چندزبانه با انتقال یادگیری زبان‌های با ساختار متفاوت[۱۳]

همانطور که مشاهده می‌کنید بهترین دقت هنگام انتقال بین زبان‌هایی است که دارای ویژگی‌های ساختاری یکسانی هستند. بنابراین در زبان‌های با ساختار متفاوت، امکان ایجاد تمایز معنایی بین جملات برای برت چندزبانه میسر نیست.

در مقاله‌ای دیگر کارتیکیان[55] و همکاران تحلیل کردند که آیا شباهت در نحوه قرارگیری کلمات جملات در دو زبان مختلف بر قابلیت یادگیری انتقالی تاثیر دارد یا خیر؟ آن‌ها برای این کار آزمایش‌هایی را برای مقایسه عملکرد متقابل زبانی با همپوشانی کلمه و بدون آن انجام دادند. با جابجایی یونیکد هر نویسه در متن ویکی‌پدیای انگلیسی با یک مقدار ثابت، یک پیکره جدید انگلیسی جعلی با نام enfake ساختند تا هیچ نوع نویسه‌ای از آن با متون ویکی‌پدیا همپوشانی نداشته باشد و انگلیسی جعلی را به‌عنوان زبانی متفاوت از انگلیسی در نظر گرفتند که دارای ویژگی‌های یکسان به جز شکل کلمه است. در فاز پیش‌آموزش مدل برت چندزبانه، با درصد جایگشت‌های مختلف ترتیب قرارگیری کلمات را تغییر دادند. آن‌ها هم ساختار زبان مبدا(enfake) و هم ساختار زبان مقصد را تغییر دادند. با توجه به این کار برت شباهت ترتیب قرار گرفتن کلمات را در موقع پیش‌آموزش نمی‌داند. اهمیت ترتیب کلمات در دو زبان مقصد و مبدا با استفاده از دو کار تشخیص موجودیت و XNLI مورد آزمون قرار گرفت. در هریک از این کارها با درصدهای مختلف، تغییر در ترتیب کلمات جملات انجام می‌شود. این درصد میزان شباهت (میزان تصادفی بودن) را کنترل می‌کند. با جایگشت‌های مختلف کلمات شامل ۰ درصد، ۲۵ درصد، ۵۰ درصد و ۱۰۰ درصد کلمات جمله، عملکرد XNLI و تشخیص موجودیت را مورد آزمایش قرار دادند و اهمیت شباهت ترتیب کلمات در دو زبان را مورد مطالعه قرار دادند. شباهت ترتیبی کلمات در دو جمله زبان مبدا و مقصد نکته‌ای است که کاملا باید در مدل برت چندزبانه مورد توجه قرار گیرد و هرچه ساختارهای جملات دو زبان مبدا و مقصد متفاوت باشد دقت پایین‌تری را به‌دست می‌آورد.

در جدول ۲ خلاصه روش‌ها و کارهای مرتبط با شباهت‌یابی معنایی بین زبانی ارائه گردیده است.

جدول ۲-خلاصه پژوهش‌های مختلف شباهت‌یابی معنایی بین زبانی

پژوهش	سال	روش	زبان
شاو [۲۸]	۲۰۱۷	استفاده از گلاو جهت استخراج بردار کلمات و در ادامه شبکه CNN	عربی-انگلیسی، اسپانیایی-انگلیسی و ترکی-انگلیسی
تیان و همکاران [۲۴]	۲۰۱۷	استفاده از ماشین ترجمه- روش ترکیبی از الگوریتم‌های GB,RF,XGB و شبکه‌های DAN,LSTM	عربی-انگلیسی، اسپانیایی-انگلیسی و ترکی-انگلیسی
وو و همکاران [۲۷]	۲۰۱۷	استفاده از فضای اطلاعات معنایی	عربی-انگلیسی، اسپانیایی-انگلیسی و ترکی-انگلیسی
تانگ و همکاران [۲۱]	۲۰۱۸	رمزگذار مشترک چندزبانه	اسپانیایی، عربی، اندونزیایی و تایلندی
چیدامبارام و همکاران [۳۴]	۲۰۱۹	تولید فضای برداری بین زبانی با استفاده از مدل مبتنی بر رمزگذار دوتایی	انگلیسی - اسپانیایی
کنو و همکاران [۳۶]	۲۰۱۹	تولید مدل بین زبانی – همترازی جملات	انگلیسی - فرانسوی
[۲۲] بریچین	۲۰۲۰	ایجاد فضاهای معنایی چندزبانه با استفاده از فرهنگ لغت‌های دوزبانه	عربی-انگلیسی
سور و ارکان [۳۸]	۲۰۲۰	استفاده از وردنت و مجموعه ترادف	انگلیسی - رومانیایی، یونانی - رومانیایی و بلغاری - ایتالیایی
بریاکو و کارپوت[۴۱]	۲۰۲۰	استفاده از برت چندزبانه و رده بند دوکلاسه	انگلیسی-فرانسوی
وانگ و همکاران [۳۹]	۲۰۲۱	استفاده از رمزنگار ممنتوم دوگانه – یادگیری متضاد	انگلیسی-چینی
دوتا[۴۲]	۲۰۲۱	استفاده از فاصله word mover جهت ایجاد فضای تعبیه بین‌زبانی	آلمانی-رومانی

همانطور که در این بخش بیان شد برت چندزبانه برای زبان‌هایی که ساختارهای متفاوتی دارند مانند فارسی و انگلیسی عملکرد مناسبی از خود نشان نمی‌دهد. اما در این پژوهش با استفاده از معماری پیشنهادی این چالش را از برت چندزبانه حل کردیم که در بخش بعدی به شرح آن خواهیم پرداخت.

۳- روش پیشنهادی

طرحی که برای سنجش شباهت معنایی بین دو جمله در دو زبان مختلف پیشنهاد می‌شود استفاده از پیکره‌های موازی جهت تغییر فضای برداری مدل‌های چندزبانه و در ادامه تنظیم‌سازی[۵۶] آن جهت شباهت‌یابی معنایی زبان منبع و زبان هدف است. روش‌های مختلفی جهت تولید فضای برداری مشترک بین کلمات وجود دارد اما برای جملات، کارهای کمتری در زمینه تولید فضای مشترک برداری جملات انجام شده است. روش پیشنهادی از پیکره موازی استفاده می‌کند و جملات در دو زبان مختلف با ساختارهای مختلف را به یک فضای برداری نگاشت می‌دهد. در این مقاله از دو مدل multilingual BERt(برت چند زبانه) و RoBERTa (برت قوی تک زبانه انگلیسی) که هر دو مبتنی بر مبدل‌ها هستند استفاده شده که از قبل آموزش داده شده‌اند و فاز تنظیم‌سازی آن در این مقاله صورت پذیرفته است. دو نکته مهم در مورد برت چند زبانه باید در نظر گرفت. نکته اولی که باید به آن اشاره کرد این است که امکان استفاده از برت چند زبانه برای زبان‌های فارسی و انگلیسی به صورت موازی وجود ندارد و فضای برداری برت چند زبانه قابلیت تشخیص شباهت معنایی بین دو جمله در دو زبان مختلف را ندارد. به همین علت از پیکره میزان و یک مدل قوی انگلیسی(Roberta) استفاده نمودیم تا هم فضای برداری مشترک بین زبان فارسی و انگلیسی ایجاد کنیم و هم اینکه غنای معنایی بردارهای خروجی از برت چند زبانه را افزایش دهیم. نکته دومی که باید به آن اشاره کرد این است باید مدل‌های مبتنی بر مبدل‌ها مانند برت چند زبانه را برای کارهای مختلف تنظیم‌سازی کرد. تنظیم‌سازی به این معناست که پارامترهای لایه آخر را طوری تغییر می‌دهیم تا برای وظیفه مشخص پردازش زبان طبیعی(مانند تشخیص موجودیت نامدار، شباهت معنایی، پرسش و پاسخ و ...) قابل استفاده باشد. شبکه برت در دو اندازه متفاوت آموزش داده شده است. برت پایه شامل ۱۲ لایه رمزگذار (که در مقاله اصلی بوک مبدل نامیده می‌شوند) و شبکه بزرگ‌تر شامل ۲۴ لایه رمزگذار تشکیل شده اند.هر رمزگذار از یک لایه خودتوجه، یک لایه متراکم[۵۷] و یک لایه dropout (نودهایی از شبکه، با احتمال p-۱ کنار گذاشته شده و نودهای دیگری با احتمال p، حفظ می‌شوند) تشکیل شده است. قبل از رمزگذارها لایه جاسازی و بعد از آن لایه‌های خروجی قرار دارد. شبکه پایه در مجموع ۱۱۰ میلیون پارامتر و شبکه بزرگ ۳۴۵ میلیون پارامتر دارد و آموزش آن ها ۴ روز طول کشیده است[۳۷].

در ادامه به جزئیات راهکار پیشنهادی و پیاده‌سازی آن بر دو زبان با ساختار متفاوت فارسی و انگلیسی می‌پردازیم:

پیکره موازی میزان [۴۴] شامل بیش از یک میلیون جفت جمله فارسی به همراه ترجمه انگلیسی آن است که برای وظایف ماشین ترجمه مورد استفاده قرار می‌گیرد. این پیکره، بزرگترین پیکره موازی فارسی-انگلیسی است که شامل بیش از ۱۲ میلیون کلمه فارسی(۱۹۸ هزار کلمه یکتا) و ۱۱ میلیون کلمه انگلیسی(۱۵۳ هزار کلمه یکتا) است. اطلاعات مربوط به ارزیابی پیکره میزان بر روی وظایف مختلف مانند ماشین ترجمه در مرجع مقاله میزان [۴۴] وجود دارد. بسیاری از مدل‌های مطرح در شبکه‌های عصبی عمیق مانند برت از پیکره‌های بزرگ استفاده نموده و جهت تولید بردار کلمات و جملات مورد استفاده قرار می‌گیرند. نکته حائز اهمیتی که در مدل‌های چندزبانه مانند برت‌چندزبانه وجود دارد، فضاهای برداری متفاوت در زبان‌های مختلف است[۱۳].

همانطور که در بخش قبل بیان شد، مدل برت چندزبانه برای زبان‌هایی که ساختارهای متفاوتی دارند(مانند زبان فارسی و انگلیسی) عملکرد مناسبی از خود نشان نمی‌دهد و همین امر باعث می‌شود که نتوان از خروجی مدل‌های چندزبانه جهت سنجش شباهت معنایی بین دوجمله در دوزبان مختلف استفاده نمود. چراکه ممکن است دو جمله دقیقا معنای یکسان داشته باشند اما در فضای برداری در دو مکان متفاوت قرار گیرند. همچنین با توجه به تمرکز اصلی این مدل‌ها بر تولید جانمایی متون به زبان‌های مختلف، جانمایی‌های تولید شده در یک زبان خاص مانند انگلیسی نیز از جانمایی‌هایی که توسط مدل تک‌زبانه مانند RoBERTa ایجاد می‌شود از کیفیت پایین‌تری برخوردار است. این نکته را می‌توان در شباهت‌یابی متون تک‌زبانه به وضوح مشاهده نمود. بنابراین می‌توان مدل چندزبانه را طوری آموزش داد که جانمایی‌های حاصل از آن به جانمایی‌های مدل تک‌زبانه خوب مانند RoBERTa نزدیک باشد. علت استفاده از این مدل نتایج خوبی است که این مدل در کارهای تک‌زبانه به‌دست آورده است. این مدل بهترین مدل تک‌زبانه انگلیسی است و جانمایی‌هایی که تولید می‌کند از بیشترین غنای معنایی برخوردار است[۴۵].

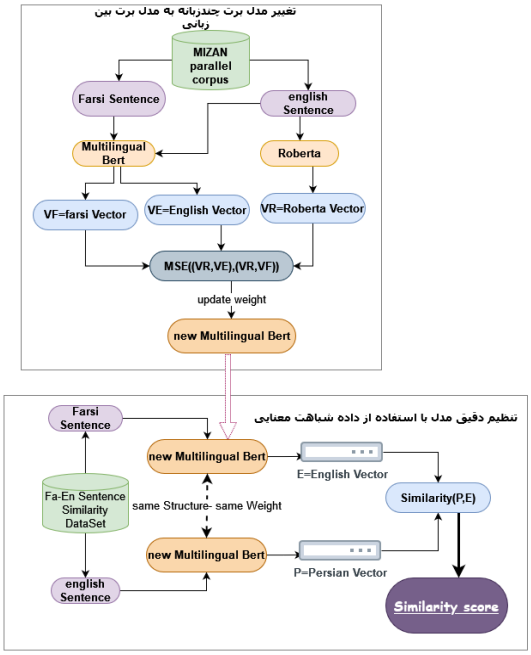
استفاده از پیکره‌های موازی برای کارهایی نظیر رده‌بندی بین زبانی متون نیز استفاده شده [۲۵] اما از آن، برای کار خاص شباهت‌یابی معنایی بین زبانی متون استفاده نشده که یکی از اهداف این مقاله همین نکته است.

با استفاده از پیکره‌های موازی می‌توان مدل‌های زبانی را ساخت که بردارهای جملات هم‌معنا به دو زبان مختلف، بسیار به هم نزدیک باشند. برای این کار می‌توان از این ایده کمک گرفت که بردار معنایی زبان مقصد باید به بردار معنایی زبان مبدا نزدیک باشد. جهت انجام این کار از پیکره موازی میزان استفاده شده و جملات انگلیسی آن به مدل RoBERTa داده شده و بردار جمله انگلیسی به‌دست آمده است. علت استفاده از مدل RoBERTa این بوده که جانمایی‌های استخراج شده از آن بیشترین کیفیت را در زبان انگلیسی دارند [۴۵]. در ادامه مدل چندزبانه برت طوری تنظیم‌سازی می‌شود که بردار جمله فارسی و جمله انگلیسی استخراج شده از آن شبیه به بردار جمله انگلیسی حاصل از مدل RoBERTa باشد. معماری طرح پیشنهادی در شکل ۲ نمایش داده شده است. یکی از معروف‌ترین و معمول‌ترین توابع زیان در تحلیل رگرسیونی، میانگین مربعات خطا[۵۸] است که به اختصار MSE نامیده می‌شود. این تابع زیان، میانگین مربعات فاصله بین مقدار پیش‌بینی و واقعی را محاسبه می‌کند. رابطه ۱ بیانگر تابع زیان مورد استفاده در این معماری است.

$$MSE = \frac{\sum[(V_r(s_{en})-V_m(s_{en}))^2+(V_r(s_{en})-V_m(s_{fa}))^2]}{n} \quad (۱)$$

در رابطه فوق منظور از V_r بردار معنایی توکن[۵۹] [cls] (cls مخفف رده بندی(classification) است و برداری است که در مدل برت بازنمایی سطح جمله را نشان می دهد) جمله انگلیسی است که با استفاده از مدل RoBERTa به‌دست آمده است. V_m نیز بردار معنایی است که از مدل برت چندزبانه حاصل شده است. S_{en} جمله انگلیسی و S_{fa} جمله فارسی است. در این رابطه n بیانگر تعداد نمونه‌ها است.

پس از آن‌که مدل چندزبانه برت با استفاده از پیکره موازی میزان آموزش داده شد(بخش بالای شکل ۲) حال از آن مدل استفاده کرده و به‌عنوان مدل اصلی در شبکه سیامی[۴۶] از آن استفاده می‌کنیم.



شکل ۲- معماری مدل شباهت معنایی بین زبانی فارسی انگلیسی پیشنهادی

شبکه‌های عصبی سیامی (که گاهی اوقات به‌عنوان یک شبکه عصبی دوقلو نیز خوانده می‌شود) یک شبکه عصبی مصنوعی است که در هنگام کار به طور پشت سر هم بر روی دو بردار ورودی مختلف، از وزن مشترک استفاده می‌کند تا بردارهای خروجی قابل مقایسه را محاسبه کند(بخش پایین شکل ۲). مولر [۴۷] با استفاده از ساختار شبکه‌های سیامی و استفاده از حافظه‌های طولانی کوتاه-مدت، مدلی با نام MaLSTM[60] را برای تشخیص میزان شباهت معنایی جفت جملات با طول متغیر ارائه داده است. خروجی مدل آن‌ها بیانگر افزایش چشم گیر دقت با استفاده از این ساختار شبکه شده است. علت استفاده از شبکه سیامی در این مقاله نیز خروجی قابل قبول آن بوده است. البته می‌توان از مدل‌های دیگر یادگیری عمیق نیز استفاده نمود لیکن با توجه به آزمایش‌های صورت پذیرفته شبکه سیامی در عین سادگی نتایج خوبی را جهت شباهت‌یابی معنایی ارائه داده است. در ادامه جملات شباهت معنایی بین زبانی فارسی – انگلیسی که از ترجمه داده محک شباهت‌یابی معنایی انگلیسی-انگلیسی به‌دست آمده، به شبکه سیامی داده شده و بردارهای مربوط به جملات فارسی و انگلیسی به دست آمدند. در پایان نیز با استفاده از معیارهای شباهت مانند شباهت کسینوسی یا فاصله اقلیدسی و منهتن بین بردار جمله فارسی و انگلیسی، میزان شباهت محاسبه می‌شود.

معماری شبکه طبق شکل ۲ است. لینک مجموعه داده و کد مقاله در دسترس عموم قرار گرفته است[۶۱].

پارامترهای یادگیری در فرآیند تنظیم‌سازی مدل به شرح زیر هستند:

تعداد دوره[۶۲]: ۴، اندازه دسته[۶۳] : ۱۶، بیشینه اندازه طول جمله : ۲۵۶، نرخ یادگیری[۶۴]: ۰.۰۰۰۰۲، بهینه‌ساز: Adam، نرخ محوشدگی(میزان کاهش نرخ یادگیری(decay))[۶۵]: ۰.۰۱.

در ادامه ارزیابی روش پیشنهادی با استفاده از داده آزمون مستخرج از ترجمه داده محک شباهت معنایی انگلیسی بیان می‌گردد.

۴- نتایج

پس از آن‌که بردار بازنمایی مربوط به هر جمله با استفاده از روش پیشنهادی در معماری شکل ۲ به‌دست آمد، میزان شباهت بین آن‌ها در فضای برداری از طریق محاسبه میزان شباهت یا عکس میزان فاصله صورت می‌پذیرد. معیارهای شباهت، معیارهایی مانند معیارهای فاصله هستند که میزان دور یا نزدیک بودن دو بردار را مشخص می‌کنند. بدیهی است که معیار شباهت با معیارهای فاصله رابطه عکس دارند و به عبارتی هر چه میزان شباهت بیشتر باشد می‌توان نتیجه گرفت فاصله‌ی دو بردار کمتر است. معیارهای متفاوتی برای محاسبه‌ی فاصله مطرح است که از جمله‌ی این معیارها می‌توان به فاصله‌ی اقلیدسی، منهتن و مینکوفسکی اشاره کرد[۴۸].

معیار شباهت کسینوسی بین دو بردار یکی از پرکاربردترین معیارها در سنجش شباهت‌یابی معنایی بین جملات است. در برخی از مقالات مربوط به تشخیص شباهت‌یابی معنایی شباهت کسینوسی را تبدیل به فاصله زاویه‌ای نموده‌اند. به‌عنوان نمونه در [۴۹] از arccos برای این کار استفاده شده است. arccos شباهت کسینوسی را به یک فاصله زاویه‌ای تبدیل می‌کند که از نابرابری مثلثی پیروی می‌کند. طبق این رویکرد، عدم وجود زاویه عملکرد بهتری را در تشخیص شباهت معنایی جملات نسبت به شباهت کسینوسی دارد. رابطه ۲ بیانگر نحوه محاسبه میزان شباهت بین دو بردار u و v با استفاده از arccos است :

$$Similarity(u,v) = -\arccos\left(\frac{u.v}{||u||||v||}\right) \qquad (۲)$$

با استفاده از معیارهای مبتنی بر فاصله مانند فاصله اقلیدسی و منهتن نیز می‌توان میزان شباهت بین دو بردار را با معکوس میزان فاصله به دست آورد. همانطور که در رابطه ۳ بیان شده، فاصله اقلیدسی کوتاه‌ترین فاصله بین دو بردار را بر طبق رابطه فیثاغورث محاسبه می‌کند. اگر x و y دو جمله با p بعد باشند، فاصله اقلیدسی بین این دو جمله به صورت رابطه ۳ است:

$$D_{euc} = \sqrt{\left(\sum_{i=1}^{p}(x_i - y_i)^2\right)} \qquad (۳)$$

اگر به جای مربع فاصله‌ی بین ابعاد جمله، از قدر مطلق فاصله بین آن‌ها استفاده شود، تابع فاصله را منهتن می‌نامند. این نام به علت تقاطع منظم خیابان‌ها در محله منهتن نیویورک انتخاب شده است. البته این فاصله گاهی به نام فاصله تاکسی یا بلوک شهری نیز نامیده می‌شود. جهت محاسبه فاصله منهتن بین دو جمله x و y با p بعداز رابطه ۴ استفاده می‌کنند:

$$D_{man} = \sum_{i=1}^{p}|x_i - y_i| \qquad (۴)$$

جهت ارزیابی خروجی سامانه‌های تشخیص شباهت معنایی از معیارهای ضریب همبستگی پیرسون[۵۰] و اسپیرمن[۵۱] استفاده می‌کنند. هدف محاسبه میزان همبستگی بین میزان شباهت تشخیص داده شده توسط سامانه با میزان شباهت واقعی آن است. نحوه محاسبه ضریب همبستگی پیرسون طبق رابطه ۵ است:

$$\Gamma_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})}} \qquad (۵)$$

هرچقدر ضریب همبستگی پیرسون به عدد یک نزدیکتر باشد مدل به‌دست آمده از دقت بالاتری برخوردار است. در ادامه نتایج مربوط به اجرای مدل بر داده‌های آزمون شباهت معنایی بین زبانی انگلیسی-فارسی بیان می‌گردد.

جهت سنجش میزان کیفیت مدل تولیدشده نیازمند داده برچسب‌خورده توسط خبره انسانی هستیم. از آنجایی که پیکره محکی برای سنجش شباهت معنایی بین زبانی فارسی انگلیسی وجود ندارد، از پیکره انگلیسی STSBenchMark [23] استفاده نموده و یک طرف جملات آن را با استفاده از ماشین ترجمه به فارسی ترجمه نمودیم تا امکان ارزیابی مدل وجود داشته باشد. این پیکره شامل ۸۶۲۸ جفت جمله به همراه میزان امتیاز شباهت معنایی در بازه ۰ تا ۵ (۰ کمترین میزان شباهت و ۵ بیشترین میزان شباهت معنایی) است که به سه قسمت داده آموزش(۷۰ درصد)، داده اعتبارسنجی(۱۵ درصد) و داده آزمون(۱۵ درصد)

تقسیم شده است. معیارهای شباهت کسینوسی، فاصله اقلیدسی و فاصله منهتن جهت محاسبه میزان شباهت بین دو بردار جمله و معیارهای پیرسون جهت سنجش میزان همبستگی بین امتیازهای مدل با امتیازهای طلایی داده آزمون مورد استفاده قرار گرفتند. نتایج حاصل از اجرای مدل به شرح جدول ۳ است.

جدول ۳- نتایج حاصل از اجرای برت چندزبانه تنظیم‌سازی شده جهت شباهت معنایی بین زبانی(مدل Raw mBERt مدل اولیه بدون تغییر- مدل New mBERt مدل تنظیم‌سازی شده با استفاده از جفت جملات فارسی -انگلیسی و داده شباهت معنایی بین زبانی)

مدل	تعداد جفت جمله (فارسی- انگلیسی)	شباهت کسینوسی(پیرسون)	فاصله اقلیدسی (پیرسون)	فاصله منهتن (پیرسون)
Raw mBERt	۰	۶۵.۰۶٪	۶۳.۶۶٪	۶۳.۶۵٪
New mBERt	۲۰۰۰۰	۶۵.۳۴٪	۶۳.۷۷٪	۶۳.۷۳٪
New mBERt	۳۵۰۰۰	۶۶.۹۸٪	۶۷.۳۱٪	۶۷.۲۱٪
New mBERt	**۱۰۰۰۰۰۰**	**۷۳.۷۷٪**	**۷۵.۳۴٪**	**۷۵.۳۷٪**

همانطور که در جدول ۳ ملاحظه می‌کنید در حالتی که فقط از مدل برت چندزبانه استفاده می‌کنیم (بدون بهینه سازی با استفاده از پیکره موازی) میزان ضریب همبستگی ۶۵ درصد به‌دست می‌آید(ردیف اول). اما با استفاده از پیکره موازی میزان ضریب همبستگی افزایش پیدا می‌کند و هرچقدر جفت جملات فارسی-انگلیسی پیکره موازی بیشتر باشد ضریب همبستگی پیرسون نیز افزایش پیدا می‌کند. به‌عنوان نمونه زمانی که از ۲۰ هزار جفت جمله فارسی انگلیسی پیکره میزان برای تولید مدل بهینه برت چندزبانه استفاده می‌کنیم(ردیف دوم) ضریب همبستگی به ۶۵.۳۴ درصد می‌رسد و با افزایش جفت جملات موازی فارسی انگلیسی از ۲۰ هزار به ۳۵ هزار ضریب همبستگی ۶۶.۹۸ به‌دست می‌آید و با افزایش آن به یک میلیون جفت جمله این ضریب به ۷۳.۷۷ درصد می‌رسد که حاکی از قابل قبول بودن روش پیشنهادی است. شایان ذکر است هرچه کیفیت ترجمه‌های داده شباهت‌یابی معنایی انگلیسی به فارسی بالاتر باشد ضریب همبستگی بیشتری نیز به‌دست می‌آید که یکی از کارهایی که در آینده می‌توان انجام داد همین نکته است.

جهت تایید کارآمدی روش پیشنهادی و آزمایش آن بر جفت زبان دیگر، مدل پیشنهادی با استفاده از ۳۰۰۰ جفت جمله عربی – انگلیسی مورد آزمایش و آزمون قرار گرفت و نتایج آن به شرح جدول ۴ است. کد و مجموعه داده در صفحه گیت‌هاب قرار داده شده است[۴۶].

جدول ۴- نتایج حاصل از اجرای روش پیشنهادی بر جفت زبان عربی- انگلیسی(معیار ضریب همبستگی پیرسون)

مدل	تعداد جفت جمله (عربی- انگلیسی)	شباهت کسینوسی	فاصله اقلیدسی	فاصله منهتن
Raw mBERt	۰	۷۶.۶۵٪	۷۳.۸۶٪	۷۳.۷۹٪
New mBERt	۳۰۰۰	**۷۸.۲۳٪**	۷۶.۷۵٪	۷۶.۸۱٪
[24] ECNU	-	۷۴.۹۳٪	-	-
BIT[27]	-	۶۹.۶۵٪	-	-

همانگونه که در جدول ۴ مشخص است روش پیشنهادی عملکرد خوبی را بر جفت زبان عربی-انگلیسی از خود نشان می‌دهد به طوریکه با استفاده از ۳۰۰۰ جفت جمله موازی می‌توان میزان ضریب همبستگی را ۲ درصد (از ۷۶.۶۵ به ۷۸.۲۳) افزایش داد. همچنین در مقایسه با سایر کارهای مرتبط روش پیشنهادی عملکرد بهتری را از خود نشان می‌دهد.

از دیگر مزایای مدل پیشنهادی عملکرد خیلی بهتر آن در محاسبه میزان شباهت معنایی بین جفت جمله فارسی و انگلیسی است به طوریکه دقیقا معنای جملات در بردارهای حاصل از جملات دو زبان وجود دارد. به‌عنوان نمونه دو جمله ساده "من یک سیب را خوردم." و "I eat an apple." را در نظر بگیرید. در مدل چندزبانه برت میزان شباهت بین این دو جمله ۰.۵۳ به‌دست آمده است در صورتی‌که در مدل چندزبانه برت بهینه شده با استفاده از روش پیشنهادی، میزان شباهت ۰.۹۶ به‌دست آمده است که بیانگر برتری آن نسبت به مدل چندزبانه برت است. این جفت جمله تنها نمونه ای از جفت جمله مورد آزمایش است و جهت درک بهتر چگونگی تاثیر مدل بیان گردیده است.

۵- نتیجه‌گیری

امروزه با پیشرفت روزافزون منابع متنی در زبان‌های مختلف ضرورت تولید مدل‌هایی که قابلیت درک همزمان بیش از یک زبان را داشته باشند بیش از پیش احساس می‌شود. یکی از مهمترین کارهایی که در پردازش زبان طبیعی مورد استفاده قرار می‌گیرد،

درک میزان معنای موجود در جملات یا عبارات است که از آن با عنوان شباهت‌یابی معنایی متون نام برده می‌شود. شباهت‌یابی معنایی متون یکی از وظایف مهم پردازش زبان طبیعی است که تحقیقات گسترده‌ای را به خود معطوف کرده است که می‌توان آن را به صورت بین زبانی نیز استفاده کرد. در این مقاله مدلی ایجاد شد که فضای برداری برت چندزبانه به فضای برداری بین زبانی فارسی-انگلیسی تغییر یافته به طوری‌که طبق آزمایش‌های انجام شده بر ترجمه انگلیسی به فارسی داده محک شباهت معنایی (STSBenchMark) نتایج حاصل بیانگر برتری روش پیشنهادی نسبت به برت چندزبانه است. استفاده از پیکره موازی فارسی-انگلیسی میزان جهت تغییر فضای برداری چندزبانه به بین‌زبانی از دیگر نوآوری‌های این پژوهش به حساب می‌آید. آزمایش‌های انجام شده حاکی از برتری روش پیشنهادی نسبت به برت چندزبانه در وظیفه شباهت معنایی بین‌زبانی متون دارد.

مراجع

[1] Manjula, D., and T. V. Geetha. "Semantic search engine." Journal of Information & Knowledge Management 3.01 (2004): 107-117.

[2] Aliguliyev, Ramiz M. "A new sentence similarity measure and sentence based extractive technique for automatic text summarization." Expert Systems with Applications 36.4 (2009): 7764-7772.

[3] De Boni, Marco, and Suresh Manandhar. "The Use of Sentence Similarity as a Semantic Relevance Metric for Question Answering." New Directions in Question Answering. (2003).

[4] Al-Anzi, Fawaz S., and Dia AbuZeina. "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing." Journal of King Saud University-Computer and Information Sciences 29.2 (2017): 189-195.

[5] Žižka, Jan, and František Dařena. "Automatic sentiment analysis using the textual pattern content similarity in natural language." International Conference on Text, Speech and Dialogue. Springer, Berlin, Heidelberg, (2010).

[6] Alzahrani, Salha M., Naomie Salim, and Ajith Abraham. "Understanding plagiarism linguistic patterns, textual features, and detection methods." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42.2 (2011): 133-149.

[7] Majumder, Goutam, et al. "Semantic textual similarity methods, tools, and applications: A survey." Computación y Sistemas 20.4 (2016): 647-665.

[8] Jaro, Matthew A. "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida." Journal of the American Statistical Association 84.406 (1989): 414-420.

[9] Winkler, William E. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." (1990).

[10] Nayantara Jeyaraj, M., & Kasthurirathna, D. (2021). MNet-Sim: A Multi-layered Semantic Similarity Network to Evaluate Sentence Similarity. arXiv e-prints, arXiv-2111.

[11] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013)

[12] Zhu, Ganggao, and Carlos A. Iglesias. "Computing semantic similarity of concepts in knowledge graphs." IEEE Transactions on Knowledge and Data Engineering 29.1 (2016): 72-85.

[13] Pires, Telmo, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. (2019).

[14] Li, Yuhua, et al. "Sentence similarity based on semantic nets and corpus statistics." IEEE transactions on knowledge and data engineering 18.8 (2006): 1138-1150.

[15] Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. "Corpus-based and knowledge-based measures of text semantic similarity." Aaai. Vol. 6. No. 2006. (2006).

[16] Agirre, Eneko, et al. "SemEval-2012 task 6: A pilot on semantic textual similarity." * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). (2012).

[17] Agirre, Eneko, et al. "* SEM 2013 shared task: Semantic textual similarity." Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity. (2013).

[18] Agirre, Eneko, et al. "SemEval-2014 task 10: Multilingual semantic textual similarity." Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). (2014).

[19] Islam, Aminul, and Diana Inkpen. "Semantic text similarity using corpus-based word similarity and string similarity." ACM Transactions on Knowledge Discovery from Data (TKDD) 2.2 (2008): 1-25.

[20] Bjerva, Johannes, and Robert Östling. "Cross-lingual learning of semantic textual similarity with multilingual word representations." 21st Nordic Conference on Computational Linguistics, NoDaLiDa, Gothenburg, Sweden,(2017).

[21] Tang, Xin, et al. "Improving multilingual semantic textual similarity with shared sentence encoder for low-resource languages." arXiv preprint arXiv:1810.08740 (2018).

[22] Brychcín, Tomáš. "Linear transformations for cross-lingual semantic textual similarity." Knowledge-Based Systems 187 (2020): 104819.

[23] Cer, Daniel, et al. "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation." (2017).

[24] Tian, Junfeng, et al. "Ecnu at SemEval-2017 task 1: Leverage kernel-based traditional nlp features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity." Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). (2017).

[25] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[26] Wieting, John, et al. "Towards universal paraphrastic sentence embeddings." arXiv preprint arXiv:1511.08198 (2015).

[27] Wu, Hao, et al. "BIT at SemEval-2017 Task 1: Using semantic information space to evaluate semantic textual similarity." Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). (2017).

[28] Shao, Yang. "Hcti at SemEval-2017 task 1: Use convolutional neural network to evaluate semantic textual similarity." Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). (2017).

[29] Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai. "Inducing crosslingual distributed representations of words." Proceedings of COLING 2012.(2012).

[30] Zou, Will Y., et al. "Bilingual word embeddings for phrase-based machine translation." Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013.

[31] Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. "Exploiting similarities among languages for machine translation." arXiv preprint arXiv:1309.4168 (2013).

[32] Gouws, Stephan, Yoshua Bengio, and Greg Corrado. "BilBOWA: fast bilingual distributed representations without word alignments." Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37.(2015).

[33] Ammar, Waleed, et al. "Massively multilingual word embeddings." arXiv preprint arXiv:1602.01925 (2016).http://arxiv.org/abs/1602.01925.

[34] Chidambaram, Muthu, et al. "Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model." Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019).(2019).

[35] Conneau, Alexis, et al. "XNLI: Evaluating Cross-lingual Sentence Representations." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.(2018).

[36] Conneau, Alexis, and Guillaume Lample. "Cross-lingual language model pretraining." Advances in Neural Information Processing Systems. (2019).

[37] Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL-HLT (1). (2019).

[38] Sever, Yiğit, and Gönenç Ercan. "Evaluating cross-lingual textual similarity on dictionary alignment problem." Language Resources and Evaluation 54.4 (2020): 1059-1078.

[39] Wang, Liang, Wei Zhao, and Jingming Liu. "Aligning Cross-lingual Sentence Representations with Dual Momentum Contrast." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

[40] He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[41] Briakou, Eleftheria, and Marine Carpuat. "Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[42] Dutta, Sourav. ""Alignment is All You Need": Analyzing Cross-Lingual Text Similarity for Domain-Specific Applications." (2021).

[43] Karthikeyan, K., et al. "Cross-Lingual Ability of Multilingual BERT: An Empirical Study." International Conference on Learning Representations. 2020.

[44] Kashefi, Omid. "MIZAN: a large persian-english parallel corpus." arXiv preprint arXiv:1801.02107 (2018).

[45] Liu, Yinhan, et al. "Roberta: A robustly optimized برت pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[46] Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[47] Mueller, Jonas, and Aditya Thyagarajan. "Siamese recurrent architectures for learning sentence similarity." Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. (2016).

[48] Singh, Archana, Avantika Yadav, and Ajay Rana. "K-means with Three different Distance Metrics." International Journal of Computer Applications 67.10 (2013).

[49] Cera, Daniel, et al. "Universal Sentence Encoder for English." EMNLP 2018 (2018): 169.

[50] Benesty, Jacob, et al. Noise reduction in speech processing. Vol. 2. Springer Science & Business Media, (2009).

[51] Benesty, Jacob, et al. Noise reduction in speech processing. Vol. 2. Springer Science & Business Media, 2009.

[52] SPEARMAN, C. " Correlation calculated from faulty data." British Journal of Psychology, 1904-1920 3.3 (1910): 271-295.

پاورقی‌ها:

1 Natural Language Processing(NLP)

2 Artificial Intelligence

3 Semantic Search

4 Summarization

5 Question-Answering System

6 Document Classification

7 Sentiment Analysis

8 Plagiarism

9 Word Embedding

10 Euclidean distance

11 Manhattan Distance

12 Cosine similarity

13 Semantic Textual Similarity BenchMark

14 character

15 Levenshtein distance

16 Hamming distance

17 Jaccard similarity coefficient

18 Deep Neural Network

19 Multilingual BERT

20 Parallel Corpus

21 Tang

22 Brychcin

23 Tian

24 Random Forest

25 Gradient Boosting

26 XGBoost

27 Projected Averaging Word Vector

28 Deep Averaging Network

29 Wu

30 semantic information space(SIS)

31 WordNet

32 Sentence Alignment

33 Shao

34 Part of Speech
35 Chidambaram
36 Conneau
37 XLU: Cross-Lingual Language Understanding
38 Swahili
39 Masked Language Model(MLM)
40 Byte Pair Encoding(BPE)
41 Sever
42 Ercan
43 SynSet
44 contrastive learning
45 Briakou
46 Carpuat
47 equivalence
48 divergence
49 Dutta
50 Pires
51 Subject-Verb-Object- SVO languages: Bulgarian, Catalan, Czech, Danish,English, Spanish, Estonian, Finnish, French, Galician, He-brew, Croatian, Indonesian, Italian, Latvian, Norwegian (Bokmaal and Nynorsk), Polish, Portuguese (European and Brazilian), Romanian, Russian, Slovak, Slovenian, Swedish, and Chinese.
52 Subject-Object-Verb- SOV Languages: Basque, Farsi, Hindi,Japanese, Korean, Marathi, Tamil, Telugu, Turkish, and Urdu.
53 Noun Adjective Order
54 Adjective Noun Order
55 Karthikeyan
56 Fine Tuning
57 Dense
58 Mean Square Error
59 Token
60 Manhattan LSTM
61https://github.com/mohammadabdous/cross-lingual-persian-english
62 Epoch
63 Batch Size
64 Learning Rate
65 Weight Decay
66 https://github.com/mohammadabdous/cross-lingual-persian-english