Digital Islamic Studies and Humanities

www.disah.inoor.ir

Monolingual and cross-lingual semantic similarity detection of Arabic texts using deep learning

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

1. PhD student in Artificial Intelligence - Faculty of Computer Engineering, Iran University of Science and Technology (IUST), Email: mohammadabdous@comp.iust.ac.ir
2. Professor at the Faculty of Computer Engineering, Iran University of Science and Technology (IUST), Email: b_minaei@iust.ac.ir

Article Info

Article type:
Research Article

Article history:
Received: 2 December 2024
Received in revised form: 28 February 2025
Accepted: 21 April 2025
Available online: 23 August 2025

Keywords:
Cross-lingual Semantic Similarity,
Arabic,
Arabic-English,
LSTM,
Deep Learning,
Siamese Network.

ABSTRACT

Semantic Textual Similarity (STS) is a crucial subfield of Natural Language Processing (NLP) that has garnered significant attention in recent years. STS aims to compute the degree of semantic similarity between two textual documents, paragraphs, or sentences, either monolingually or cross-lingually. In this paper, we focus on calculating the semantic similarity between two sentences in Arabic and Arabic-English cross-lingual pairs. Given the prevalence of Arabic texts in Islamic literature, this research has numerous practical applications. The semantic similarity between two sentences can be determined using their semantic vectors. To compute this similarity, we first need to represent each sentence as a vector .

In this study, word vectors were extracted using pre-trained embeddings on Arabic texts from Twitter and Wikipedia, employing two well-known word embedding techniques: CBOW and Skip-Gram. Additionally, transformer-based models such as paraphrase-xlm-roberta were utilized for cross-lingual semantic similarity calculation between Arabic and English. To evaluate and train the model, we used data from the Semantic Textual Similarity Conference of 2017, which includes Arabic-Arabic and Arabic-English sentence pairs. A deep neural network model, specifically a Siamese network with an LSTM layer, was trained. LSTM enables the network to learn long-term dependencies. Siamese networks, despite their simplicity, yield satisfactory results, while transformer-based models demonstrate strong cross-lingual learning capabilities.

In the final layer of the network, the cosine similarity between the vectors of the two input sentences is used to determine their degree of similarity. The results indicate that the proposed method achieves a Pearson correlation of 83.4% for Arabic-Arabic sentence pairs and 82% for Arabic-English pairs, outperforming other existing approaches.

Cite this article: Abdous, M., & Behrouz Minaei. (2025). Monolingual and cross-lingual semantic similarity detection of Arabic texts using deep learning. *Digital Islamic Studies and Humanities*, 1 (1), 103-122. https://doi.org/10.22034/disah.2024.716150

 Publisher: Research Center for Digital Islamic Studies and Humanities (RCDISAH).
DOI: https://doi.org/10.22034/disah.2024.716150

علوم اسلامی و انسانی دیجیتال

www.disah.inoor.ir

شباهت‌یابی معنایی تک زبانه و بین زبانی متون عربی با استفاده از یادگیری عمیق

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

۱. دانشجوی دکتری هوش مصنوعی – دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران: mohammadabdous@comp.iust.ac.ir

۲. دانشیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران: b_minaei@iust.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

تاریخ دریافت: ۱۴۰۳/۰۹/۱۲
تاریخ بازنگری: ۱۴۰۳/۱۲/۱۰
تاریخ پذیرش: ۱۴۰۴/۰۲/۰۱
تاریخ انتشار: ۱۴۰۴/۰۶/۰۱

کلیدواژه‌ها:
شباهت‌یابی معنایی بین‌زبانی، عربی، عربی – انگلیسی، LSTM، یادگیری عمیق، شبکه سیامی.

چکیده

شباهت‌یابی معنایی متون (STS) یکی از زیرشاخه‌های مهم پردازش زبان طبیعی است که در چند سال اخیر توجه بسیاری را به خود جلب کرده است. هدف در شباهت‌یابی معنایی، محاسبه میزان شباهت معنایی بین دو سند متنی، پاراگراف یا جمله است که به‌صورت تک‌زبانه و بین‌زبانی مطرح می‌شود. در این مقاله، شباهت‌یابی معنایی بین جملات زبان عربی و جملات بین‌زبانی عربی- انگلیسی بررسی شده است؛ با توجه به کاربرد گسترده متون اسلامی به زبان عربی، این پژوهش از اهمیت بسیاری برخوردار است. برای محاسبه شباهت معنایی بین دو جمله، از بردارهای معنایی آن‌ها استفاده می‌شود. در این تحقیق، با به‌کارگیری بردارهای از پیش آموزش‌داده‌شده بر روی متون عربی موجود در توییتر و ویکی‌پدیا و به‌کمک دو روش CBOW و Skip-Gram که از معروف‌ترین روش‌های آموزش تعبیه کلمات هستند، بردارهای کلمات استخراج شده است. همچنین، از مدل‌های مبتنی بر مبدل‌ها مانند paraphrase-xlm-roberta نیز برای محاسبه شباهت معنایی بین‌زبانی عربی-انگلیسی استفاده شده است. برای ارزیابی و آموزش مدل، از داده‌های کنفرانس شباهت‌یابی معنایی سال ۲۰۱۷ شامل جفت جمله‌های عربی و جفت جمله‌های عربی-انگلیسی استفاده کرده و یک شبکه عصبی عمیق به نام شبکه سیامی با لایه LSTM را آموزش دادیم. LSTM امکان یادگیری وابستگی‌های بلندمدت را در شبکه فراهم می‌کند. شبکه‌های سیامی در عین سادگی نتایج قابل‌قبولی ارائه می‌دهند و مدل‌های مبتنی بر مبدل نیز قابلیت یادگیری بین‌زبانی دارند. در لایه نهایی شبکه، شباهت کسینوسی بین بردارهای متعلق به دو جمله ورودی، میزان شباهت آن‌ها را محاسبه می‌کند. نتایج نشان می‌دهد که روش پیشنهادی به همبستگی پیرسون ۸۳٫۴ درصد برای جفت جمله‌های عربی-عربی و ۸۲ درصد برای جفت جمله‌های عربی- انگلیسی دست یافته است، که عملکرد بهتری نسبت به سایر روش‌های موجود دارد.

استناد: عبدوس، محمد؛ و مینایی بیدگلی، بهروز (۱۴۰۴). شباهت‌یابی معنایی تک زبانه و بین زبانی متون عربی با استفاده از یادگیری عمیق. علوم انسانی و اسلامی دیجیتال، ۱ (۱)، ۱۰۳-۱۲۲. https://doi.org/10.22034/disah.2024.716150

ناشر: پژوهشکده علوم اسلامی و انسانی دیجیتال (مرکز تحقیقات کامپیوتری علوم اسلامی نور).

مقدمه

سنجش تشابه معنایی متون[1] بین کلمات، جملات، پاراگراف‌ها و اسناد نقش مهمی در علوم کامپیوتر و زبان‌شناسی محاسباتی ایفا می‌کند و کاربردهای گسترده‌ای در حل مسائل مختلف پردازش زبان طبیعی[2] دارد. در ترجمه ماشینی، برای مقایسه معنایی جمله تولیدشده توسط سیستم با جمله در زبان اصلی، از این روش استفاده می‌شود (کوملس[3] و آتسریاز[4]، ۲۰۱۹). در استخراج اطلاعات نیز شباهت‌یابی معنایی به منظور یافتن روابط بین متون منابع مختلف کاربرد دارد (لوبانی و همکاران، ۲۰۱۹).

یکی از کاربردهای شناسایی متون مشابه، تشخیص متون ویرایش‌شده است (رول[5] و همکاران، ۲۰۲۰). همچنین، شباهت‌یابی معنایی در کشف تقلب در کدهای برنامه‌نویسی، سیستم‌های پرسش و پاسخ و ابهام‌زدایی واژگان نیز به کار می‌رود (داس[6] و همکاران، ۲۰۲۲).

محاسبه شباهت بین متون کوتاه برای اولین بار در سال ۲۰۰۶ گزارش شد (لی و همکاران، ۲۰۰۶؛ میهالسیا[7] و همکاران، ۲۰۰۶). پس از آن، از سال ۲۰۱۲ به بعد، هدف شباهت‌یابی معنایی تنها یافتن شباهت نبود، بلکه محاسبه دقیق میزان شباهت برای هر جفت متن (جمله) با عددی بین ۰ تا ۵ مورد توجه قرار گرفت (اگیرا[8] و همکاران، ۲۰۱۲-۲۰۱۳-۲۰۱۴). در این رویکرد، عدد ۰ بیانگر عدم ارتباط معنایی و عدد ۵ نشان‌دهنده برابری کامل معنایی بین دو متن است.

ایده‌های اولیه برای شناسایی شباهت معنایی بین دو جمله، بر اساس هم‌ترازی معنایی بین کلمات دو جمله و جمع جبری شباهت‌های کلمات بود (اسلام و همکاران، ۲۰۰۸). امروزه، بیشتر پژوهش‌ها بر بازنمایی معنایی جملات با استفاده از یادگیری عمیق تمرکز دارند. در این روش‌ها، جملات به بردارهای عددی با ابعاد مختلف تبدیل می‌شوند که حامل معانی کلمات در فضای برداری هستند. در این فضا، کلماتی که به یکدیگر نزدیک‌ترند، از نظر معنایی نیز به هم شباهت بیشتری دارند.

1. Semantic Textual Similarity.
2. Natural Language Processing.
3. Comelles.
4. Atserias.
5. Roul.
6. Das.
7. Mihalcea.
8. Agirre.

تولید بردارهای معنایی کلمات با استفاده از پیکره‌های بزرگ متنی انجام می‌شود. در زبان انگلیسی، به دلیل گستردگی کشورهای انگلیسی‌زبان و کاربرد جهانی آن، پژوهش‌های بیشتری در این زمینه صورت گرفته است. اما در زبان‌هایی با منابع و پیکره‌های[1] محدودتر، مانند عربی و فارسی، تحقیقات کمتری انجام شده است. یکی از بهترین پیکره‌های موجود که به‌صورت آزاد برای زبان‌های مختلف قابل بهره‌برداری است، ویکی‌پدیا است که امروزه تقریباً برای تمامی زبان‌های دنیا در دسترس است.

در این پژوهش، با استفاده از بردارهای از پیش آموزش‌داده‌شده بر روی متون شبکه اجتماعی توییتر عربی، سیستمی برای محاسبه میزان شباهت معنایی بین دو جمله طراحی کرده‌ایم که نتیجه آن عددی بین ۰ تا ۵ است. این کار با به‌کارگیری یادگیری عمیق و به‌طور خاص، استفاده از شبکه سیامی[2] انجام شده است. برای بازنمایی بردار جمله از LSTM استفاده شده و در نهایت، با محاسبه شباهت کسینوسی[3] بین دو بردار، میزان شباهت معنایی بین دو جمله به‌دست می‌آید؛ هر چه این عدد بزرگ‌تر باشد، نشان‌دهنده شباهت معنایی بیشتر است. همچنین با استفاده از داده‌های بین‌زبانی عربی-انگلیسی، مدلparaphrase-xlm-r-multilingual را آموزش داده‌ایم.

ساختار مقاله بدین ترتیب است: در بخش دوم، پژوهش‌های مرتبط در زمینه شباهت‌یابی معنایی به‌اختصار بررسی می‌شود، سپس روش پیشنهادی توضیح داده می‌شود و در بخش چهارم، نتایج و نتیجه‌گیری مقاله ارائه می‌شود.

الف. کارهای مرتبط

درباره شباهت‌یابی معنایی، پژوهش‌های بسیاری انجام شده است که در اینجا به برخی از مهم‌ترین آن‌ها اشاره می‌کنیم. اغلب این پژوهش‌ها بر روی زبان انگلیسی متمرکز بوده و به همین دلیل، این بخش بیشتر بر پژوهش‌های زبان انگلیسی تمرکز دارد و در پایان نیز به چند پژوهش در زمینه زبان عربی و شباهت‌یابی بین‌زبانی عربی-انگلیسی اشاره می‌شود.

1. Corpus.
2. Siamese.
3. Cosine Similarity.

یکی از مهم‌ترین کنفرانس‌ها در حوزه شباهت‌یابی معنایی، کنفرانس‌های SEM است که سیستم‌های مختلف در آن به رقابت می‌پردازند. ابتدا به بررسی بهترین سیستم‌های هر دوره از این کنفرانس می‌پردازیم. بار[۱] و همکاران (۲۰۱۲) در SEM2012 رتبه اول را به‌دست آوردند و از مدل رگرسیون خطی لگاریتمی برای محاسبه شباهت معنایی استفاده کردند. ویژگی‌های ورودی مدل آن‌ها شامل بزرگ‌ترین زیررشته و توالی مشترک،[۲] اشتراک بین ان-گرام‌ها و ضریب جاکارد[۳] بود. همچنین از نمایش گرافی کلمات و شبکه واژگان برای محاسبه ارتباط معنایی بهره بردند.

هان[۴] و همکاران (۲۰۱۳) در SEM2013 سیستمی با بالاترین معیار همبستگی پیرسون به نام Ebiquity-Coreارائه کردند. آن‌ها شباهت معنایی مبتنی بر LSA را با شبکه واژگان[۵] ترکیب کردند و قواعدی برای افزایش شباهت بین کلمات مرتبط نظیر مترادف‌ها و واژه‌های دارای ارتباط مستقیم (مانند پزشک و بیمارستان) در نظر گرفتند.

سلطان[۶] و همکاران (۲۰۱۴ و ۲۰۱۵) در SEM2014 و SEM2015 سیستمی با نام DLS@CUمعرفی کردند که با استفاده از هم‌ترازی معنایی بین کلمات عملکرد بسیار خوبی داشت. در این سیستم، دو کلمه هم‌تراز محسوب می‌شوند اگر از لحاظ معنایی شبیه به هم باشند یا در بافتارهایی با شباهت زیاد قرار گیرند.

ریچالسکا[۷] و همکاران (۲۰۱۶) در SEM2016 سیستمی به نام Samsung-Ensemble را ارائه دادند که بهترین عملکرد را بر اساس معیار همبستگی پیرسون داشت. آن‌ها از رده‌بندهای مختلف مبتنی بر رگرسیون بردار پشتیبان خطی[۸] (LSVR) و شبکه عصبی GRU بهره بردند.

---

1. Bar.
2. Longest Common Substring
3. jaccard coefficient.
4. Han.
5. WordNet.
6. Sultan.
7. Rychalska.
8. Linear Support Vector Regression.

مولر و تیاگاراجان (۲۰۱۶) از شبکه عصبی عمیق با ساختار سیامی برای شناسایی شباهت جملات استفاده کردند. آن‌ها ازLSTM یک‌طرفه برای تبدیل بردار کلمات به بازنمایی معنایی جمله بهره بردند و در لایه نهایی از فاصله منهتن[1] برای محاسبه میزان شباهت استفاده کردند.

از سال ۲۰۱۷ به بعد، تمرکز کنفرانس SEM بیشتر به محاسبه شباهت بین جملات در زبان‌های مختلف معطوف شد. به‌عنوان مثال، درSEM2017 ، شباهت یک جمله عربی با یک جمله انگلیسی محاسبه می‌شد (کر و همکاران، ۲۰۱۷).

در زبان عربی، کنفرانس SEM2017 اولین پیکره شباهت‌یابی زبان عربی را با ۱۰۸۱ جفت جمله عربی تولید کرد. در این کنفرانس، یکی از آیتم‌های رقابتی محاسبه شباهت معنایی بین جفت جملات عربی بود. مدل BIT (وو و همکاران، ۲۰۱۷) با شباهت ۷۵٫۴۳ بهترین عملکرد را داشت و از ترکیب یک سیستم بدون نظارت و دو سیستم باناظر استفاده کرد. سیستم‌های آن‌ها عمدتاً به فضای اطلاعات معنایی (SIS) وابسته بودند که بر اساس طبقه‌بندی سلسله‌مراتبی معنایی در شبکه واژگان، محتوای غیرهم‌پوشان جملات را محاسبه می‌کرد و بهترین عملکرد را در مجموعه داده عربی-عربی به‌دست آوردند.

اسچواب[2] و همکاران (۲۰۱۷) از بازنمایی کلمات در یک فضای چندبعدی برای نمایش ویژگی‌های معنایی و نحوی بهره بردند. آن‌ها با استفاده از ویژگی‌هایی مانند وزن IDF[3] و برچسب اجزای سخن،[4] کلمات توصیفی مهم هر جمله را شناسایی کردند. برای محاسبه IDF، هر جمله به‌عنوان یک سند در نظر گرفته شد و وزن‌دهی کلمات بر همین مبنا انجام شد. ارزیابی سیستم آن‌ها بر روی ۷۵۰ جفت جمله عربی از پیکره توضیحات فیلم،[5] که به زبان عربی ترجمه شده بود، نشان داد که استفاده از برچسب اجزای سخن و وزن‌دهی کلمات، همبستگی پیرسون را ۷ درصد افزایش داده و به ۷۹٫۶۹ درصد رسانده است.

1. Manhattan Distance.
2. Schwab.
3. inverse document frequency
4. Part Of Speech Tag.
5. MSR-Video.

بیجروا[1] و همکاران (۲۰۱۷) سیستمی برای شباهت‌یابی معنایی در متون چندزبانه ارائه کردند. ایده اصلی آن‌ها بر این بود که کلمات با معنای مشابه، بازنمایی‌های نزدیکی در فضای برداری داشته باشند. آن‌ها با بهره‌گیری از داده‌های موازی در زبان‌های انگلیسی، اسپانیایی و عربی، مدل‌های مختلفی برای شباهت‌یابی معنایی چندزبانه ایجاد کردند و با آزمایش بر پیکره عربی- عربی به همبستگی ۷۱ درصد دست یافتند.

تیان[2] و همکاران (۲۰۱۷) با ترکیب مدل‌های مبتنی بر ویژگی و شبکه‌های عصبی عمیق، بهترین سیستم را از نظر معیار همبستگی پیرسون در SEM2017 ارائه دادند. مدل‌های مبتنی بر ویژگی شامل Random Forest (RF)، گرادیان تقویتی (GB) و XGBoost (XGB) بودند، و مدل‌های مبتنی بر شبکه‌های عصبی عمیق شامل میانگین تعبیه کلمات، تعبیه کلمات ترازشده،[3] شبکه DAN[4] و شبکهLSTM . آن‌ها با ترکیب این مدل‌ها سیستمی ساختند که برای تشخیص شباهت جملات استفاده می‌شود و بر روی زبان عربی به همبستگی ۷۴/۴ درصد دست یافتند.

بریچین[5] (۲۰۲۰) ایده‌ای مطرح کرد که در آن فضاهای معنایی چندزبانه با استفاده از فرهنگ لغت‌های دوزبانه در یک فضای مشترک قرار می‌گیرند. آن‌ها سیستمی مبتنی بر تکنیک‌های بدون نظارت برای شباهت‌یابی جملات ایجاد کرده و نشان دادند که می‌توان با وزن‌دهی به کلمات، فضاهای معنایی مشترک را بهبود بخشید. نتایج آن‌ها برای جفت جملات عربی-عربی همبستگی ۶۹ درصد را نشان داد.

همچنین، تیان و همکاران (۲۰۱۷) با روشی به نام ECNU که ترکیبی از روش‌های سنتی پردازش زبان طبیعی و شبکه‌های عصبی است، ویژگی‌هایی را استخراج کرده و با میانگین‌گیری از امتیازهای حاصل از این روش‌ها، امتیاز شباهت نهایی را به‌دست آوردند. این روش توانست بهترین امتیاز را در شباهت‌یابی جملات بین‌زبانی برای جفت زبان‌های عربی- انگلیسی در مجموعه داده‌های SemEval17 کسب کند.

1. Bjerva.
2. Tian.
3. Projected word embedding.
4. Deep Averaging Network.
5. Brychcin.

در این فصل، کارهای مرتبط در زمینه شباهت‌یابی معنایی برای زبان عربی-عربی معرفی شد. روش‌های مختلفی برای محاسبه شباهت جملات عربی وجود دارد که غالباً پیچیده هستند؛ با این حال، روش پیشنهادی این پژوهش به دلیل سادگی، نتایج بهتری نسبت به سایر روش‌ها به‌دست آورده است.

ب. روش شباهت‌یابی معنایی پیشنهادی

در این بخش معماری سیستم شباهت‌یابی معنایی پیشنهادی برای زبان عربی و زبان‌های بین‌زبانی (عربی- انگلیسی) تشریح می‌شود. رویکرد پیشنهادی برای شناسایی شباهت معنایی بین جملات، استفاده از شبکه عصبی عمیق است. اولین گام در شناسایی شباهت معنایی، تبدیل هر جمله به یک بردار است. این بردارهای کلمات معمولاً از پیکره‌های بزرگ متنی مانند ویکی‌پدیا، توییتر یا متون کتاب‌ها و روزنامه‌ها استخراج می‌شوند. برای استخراج بردار کلمات عربی، از مقاله‌ای نوشته سلیمان و همکاران (۲۰۱۷) استفاده کردیم که مدل‌های از پیش آموزش‌داده‌شده زبان عربی را ارائه کرده‌اند. آن‌ها نام این مدل را araVec گذاشته‌اند که مخففArabic vector است. در پژوهش خود، آن‌ها مدل‌های متعددی برای استخراج بردار کلمات با استفاده از دو روش CBOW وSkipGram بر روی متون عربی توییتر و ویکی‌پدیا ایجاد کرده‌اند. در اینجا توضیح مختصری پیرامون هر یک از این مدل‌ها ارائه می‌کنیم. آن‌ها برای هرکدام از دو روش مذکور، مدل‌های مختلفی از تعبیه کلمات با ابعاد ۱۰۰ و ۳۰۰ تولید کرده‌اند؛ اما در این پژوهش، با توجه به محدودیت سخت‌افزاری، تنها از مدل ۱۰۰ بعدی استفاده شده است.

برای شناسایی شباهت معنایی بین‌زبانی (عربی-انگلیسی)، از یک مدل مبتنی بر مبدل استفاده شده است. مدل مورد استفاده برای تشخیص میزان شباهت معنایی بین جملات عربی و انگلیسی، مدل paraphrase-xlm-r-multilingual است.

در این پژوهش، از شبکه سیامی برای محاسبه میزان شباهت معنایی بین جملات بهره گرفته‌ایم که در پژوهش‌های پیشین برای زبان عربی از این شبکه استفاده نشده است. این موضوع یکی از نوآوری‌های مهم این پژوهش به‌شمار می‌آید. همچنین، در بخش نتایج نشان خواهیم داد که این رویکرد عملکرد بهتری نسبت به روش‌های پیشین دارد.

۱. معرفی داده

از مهم‌ترین بخش‌های هر سیستم شباهت‌یابی معنایی، وجود مجموعه‌داده‌ای است که هم برای آموزش مدل و هم برای ارزیابی آن قابل‌استفاده باشد. در زبان عربی، داده‌های موجود برای شباهت‌یابی معنایی از کنفرانس SEM2017 گرفته شده‌اند.[1] این مجموعه‌داده شامل ۱۰۸۱ جفت جمله عربی است که میزان شباهت هر جفت جمله به‌صورت عددی بین ۰ تا ۵ تعیین شده است. این مجموعه‌داده از سه منبع مختلف به‌دست آمده است: ۵۱۰ جفت جمله از پیکره MRPC ، ۳۶۸ جفت جمله از پیکره MRVDC ، و ۲۰۳ جفت جمله از پیکره[2] ترجمه ماشینی.

برای شباهت‌یابی معنایی بین‌زبانی عربی- انگلیسی، از ۱۱۰۰ جفت جمله عربی-انگلیسی استفاده شده است.

در هنگام استفاده از این داده‌ها، برخی از مراحل پیش‌پردازش متنی انجام می‌شود؛ مانند حذف فاصله‌های اضافی، یکسان‌سازی حروف «ی»، «ک»، و «ة» و حذف علامت‌های فتحه، ضمه، کسره، و تشدید. در جداول (۱ و ۲) بخشی از این مجموعه‌داده‌ها را مشاهده می‌کنید.

جدول (۱): نمونه‌ای از جملات مجموعه داده

جمله اول	جمله دوم	میزان شباهت
إمرأة صغيرة تعزف الناي.	إمرأة تغسل الجزء العلوي من المجمد.	۰
إمرأة تقص زهورا.	رجل يحمل زهور عباد الشمس الضخمة.	۱
إمرأة تقطّع حبّة فلفل كبير.	إمرأة تقطّع فلفلا أخضرا.	۳٫٤
المرأة تضع المكياج للرجل.	إمرأة ترسم وجه رجل.	۲٫۸
في آخر مرة تم فيها إجراء المسح، عام ۱۹۹۵، تم تطابق تلك الأرقام.	في المسح الأخير من عام ۱۹۹۵، كانت تلك الأرقام متساوية.	۵

1. alt.qcri.org/semeval2017/task1.
2. Microsoft Research Paraphrase Corpus.

جدول (۲): نمونه‌ای از جفت جملات عربی ـ انگلیسی

جمله عربی	جمله انگلیسی	امتیاز
إن حق الحكومة في أن تنبذ اعتباطيا دستورها هو التعريف الإستبداد المتميز.	The right of a government arbitrarily to set aside its own constitution is the defining characteristic of a tyranny.	٤٫٨
أجل ملك مقاطعة المحكمة العليا القاضي تشارلز ملتر عقد المحاكمة، حتى يوم الإثنين.	After Freitas' opening statement, King County Superior Court Judge Charles Mertel recessed trial until after the Thanksgiving weekend.	٣
بريئة أم لا، "فقد أتلفت البضائع "هذا ما قاله هوارد دافيتوتز، رئيس شركة دافيدوتزوشركاه، وهي شركة استشارية الوطنية متجزئة في نيويورك .	Get it all out," says Howard Davidowitz, chairman of Davidowitz & Associates, a national retail consulting firm based in New York City.	١٫٤
في عام ٢٠٠٦، ستعطي شركة آجا باسيفيك تقريرا عن نمو بنسبة ٧٫٩ في المئة ليصل إلى ٨١٨ بليون دولار .	In ٢٠٠٦, the group says the market will rebound ٢٩٫٦ percent to $٢١٫٣ billion in sales.	٢
من المتوقع تحقيق أرباح للعمليات التشغيلية في نيسان في حدود ٦٠ إلى ٨٠ سنتا للسهم الواحد.	Kodak expects earnings of ٥ cents to ٢٥ cents a share in the quarter.	١

آخرین رقابت بین‌المللی برای ارزیابی روش‌های مختلف شباهت‌یابی معنایی جملات، Semeval2017است، که مجموعه داده‌های متنوعی را برای هر جفت زبان فراهم می‌کند. بیشتر روش‌های شباهت‌یابی جملات از این مجموعه داده‌ها برای ارزیابی استفاده کرده و برخی نیز از مجموعه‌های داده جداگانه آن برای آموزش مدل‌ها بهره می‌برند.

برای جفت زبان عربی-انگلیسی از مجموعه داده Semeval 2017 استفاده کرده‌ایم. این مجموعه شامل ۱۹۱۲ جفت جمله برای آموزش و ۲۵۰ جفت جمله برای ارزیابی است. تعداد جفت‌جملات موجود در پیکره عربی-انگلیسی برای آموزش مدل‌ها در شکل (۱) نشان داده شده است.

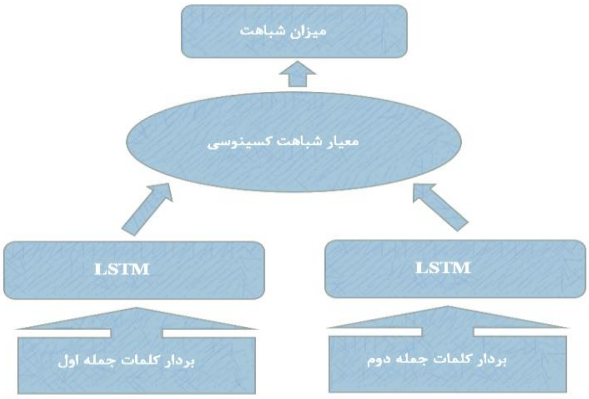
شکل (۱):آمار جفت جملات عربی-انگلیسی در پیکره SemEval2017 جهت آموزش مدل‌ها

Year	Data set	Pairs	Source
2017	Trial	23	Mixed STS 2016
2017	MSRpar	1020	newswire
2017	MSRvid	736	videos
2017	SMTeuroparl	406	WMT eval.

۲. معماری مدل پیشنهادی

شبکه‌های سیامی[1] به‌طور گسترده‌ای در انواع مدل‌های یادگیری عمیق به کار می‌روند. در این مقاله، از شبکه سیامی به همراه سلول‌های LSTM برای محاسبه میزان شباهت بین دو جمله عربی استفاده شده است. یکی از مراحل اصلی در شناسایی میزان شباهت بین جملات، تبدیل جملات از فضای متنی به فضای برداری است. در این فرآیند، ابتدا کلمات تشکیل‌دهنده جملات به بردارهای عددی تبدیل می‌شوند و سپس با استفاده از روش‌های مختلف، مانند میانگین‌گیری از بردارهای کلمات جمله، بردار نهایی جمله به‌دست می‌آید. معماری مدل پیشنهادی برای تشخیص میزان شباهت معنایی بین دو جمله عربی در شکل (۲) نمایش داده شده است.

شکل (۲): معماری مدل پیشنهادی



در ادامه، توضیحاتی پیرامون نحوه ایجاد بردار کلمات ارائه داده می‌شود. دو روش اصلی برای یادگیری تعبیه کلمات وجود دارد که هر دو به دانش محتوایی وابسته هستند:

۳. مدل فضای برداری مبتنی بر شمارش[2]

مدل‌های فضای برداری مبتنی بر شمارش با فرض اینکه کلمات در یک محتوای مشابه از نظر معنایی[3] مرتبط یا مشابه هستند، به شدت وابسته به نرخ رخداد کلمات (فرکانس کلمات) و ماتریس

1. Siamese.
2. Count based Vector Space Model.
3. semantic meanings.

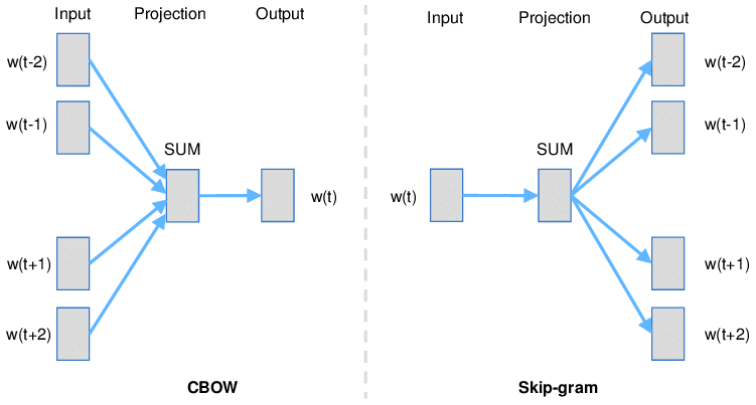
هم‌رخدادی‌اند.[1] این مدل‌ها آمارهای مبتنی بر شمارش مانند هم‌رخدادی بین کلمات همسایه را به یک بردار کلمه کوچک و متراکم نگاشت می‌کنند. تحلیل مولفه‌های اصلی،[2] مدل‌های موضوعی[3] و مدل‌های زبانی احتمالاتی عصبی،[4] همگی نمونه‌های خوبی از این دسته به شمار می‌روند (شهمیرزادی و همکاران، ۲۰۱۹).

روش‌های مبتنی بر محتوا برخلاف روش‌های مبتنی بر شمارش عمل می‌کنند. در این روش‌ها، میزان احتمال حضور یک کلمه با توجه به کلمات همسایه آن (سمت چپ و راست) محاسبه می‌شود. بهترین بازنمایی برداری از هر کلمه در زمان آموزش مدل تولید می‌شود. در این بخش به معرفی دو نوع از این مدل‌ها خواهیم پرداخت:

۴. مدل‌های مبتنی بر محتوا[5]

مدل‌های CBOW و Skip-gram از دیگر مدل‌هایی هستند که برای یادگیری بردارهای کلمات مورد استفاده قرار می‌گیرند. این مدل‌ها توسط میکولوف و همکاران (۲۰۱۳) معرفی شده‌اند. معماری این مدل‌ها در شکل (۲) قابل مشاهده است:

شکل (۳): مدل **CBOW** و **Skip-gram** (میکولوف و همکاران، ۲۰۱۷)



1. co-occurrence matrix.
2. Principal Component Analysis.
3. topic models.
4. neural probabilistic language models.
5. Context based Vector Space Model.

همان‌طور که در شکل (۲) مشاهده می‌کنید، در مدلCBOW، هدف پیش‌بینی کلمه‌ای است که در وسط پنجره متحرک قرار دارد، با توجه به کلمات قبل و بعد از آن. در این مدل، W(t) کلمه هدف در زمان t است. فرض کنید یک پنجره متحرک با اندازه ثابت در امتداد یک جمله حرکت می‌کند. کلمه‌ای که در وسط این پنجره قرار می‌گیرد "هدف" است، و کلماتی که در سمت چپ و راست آن در داخل پنجره قرار دارند، کلمات محتوایی به‌شمار می‌روند.

در مدلSkip-gram ، هدف پیش‌بینی احتمال اینکه آیا کلمه داده‌شده به‌ازای یک "هدف"، کلمه محتوایی است یا خیر. در واقع، در این مدل، کلمات سمت راست و چپ کلمه هدف پیش‌بینی می‌شوند. از مزایای مدل Skip-gram این است که با داده‌های آموزش کمتر نیز به‌خوبی عمل می‌کند و حتی برای کلمات و عبارات کمیاب نیز بازنمایی خوبی ارائه می‌دهد. در مقابل، در مدل CBOW فرآیند آموزش سریع‌تر از Skip-gram است و دقت بالاتری برای کلمات پرکاربرد ارائه می‌دهد.

در این مقاله از هر دو مدل برای بازنمایی برداری کلمات عربی استفاده شده است. پس از به‌دست‌آوردن بازنمایی برداری کلمات، نیاز داریم تا کل جمله را به‌صورت یک بردار نمایش دهیم تا بتوانیم با استفاده از شباهت کسینوسی میزان شباهت بین دو جمله را استخراج کنیم. برای بازنمایی کل جمله از کلمات جمله، از LSTM استفاده کردیم. استفاده از LSTM این امکان را فراهم می‌آورد که ماشین یادگیرنده توالی‌های کلمات را به‌خوبی یاد بگیرد و بردار نهایی که از آخرین سلول LSTM استخراج می‌شود، شامل بردار معنایی کل جمله باشد. در ادامه به توضیحاتی پیرامون LSTM پرداخته خواهد شد.

۵. معرفی LSTM

شبکه‌های (LSTM) خلاصه شده از عبارت "Long Short Term Memory" نوع خاصی از شبکه‌های عصبی بازگشتی هستند که توانایی یادگیری وابستگی‌های بلندمدت را دارند. این شبکه‌ها برای اولین بار توسط هچریتر در سال ۱۹۹۷ معرفی شدند (هچریتر[1] و همکاران، ۱۹۹۷). هدف اصلی طراحی شبکه‌های LSTM، حل مشکل وابستگی بلندمدت در داده‌های دنباله‌ای است. در واقع، به یاد سپاری اطلاعات برای بازه‌های زمانی بلندمدت، رفتار پیش‌فرض و عادی شبکه‌های LSTMاست. ساختار این شبکه‌ها به‌گونه‌ای طراحی شده است که اطلاعات دوردست را به‌خوبی یاد می‌گیرند، و این ویژگی در طراحی آن‌ها نهفته است.

1. Hochreiter.

تمام شبکه‌های عصبی بازگشتی به شکل دنباله‌ای (زنجیره‌ای) تکرار شونده از واحدهای شبکه عصبی ساخته می‌شوند. در شبکه‌های عصبی بازگشتی استاندارد، این واحدهای تکرار شونده ساختار ساده‌ای دارند، به‌طوری که معمولاً تنها شامل یک لایه تانژانتِ هایپربولیک (tanh) هستند. در مقابل، شبکه‌های LSTM نیز ساختار دنباله‌ای دارند، اما ماژول تکرار شونده آن‌ها ساختاری پیچیده‌تر است. به جای داشتن یک لایه شبکه عصبی، این ماژول‌ها شامل چهار لایه هستند که طبق یک ساختار ویژه با یکدیگر در تعامل هستند و در نهایت منجر به ایجاد خروجی نهایی می‌گردند.

در این مقاله، پس از به‌دست‌آوردن بردارهای کلمات با استفاده از مدل‌های مبتنی بر محتوا، این بردارها به شبکه LSTM داده می‌شود تا بردار نهایی جمله تولید گردد. سپس، با استفاده از شباهت کسینوسی بین بردارهای نهایی جمله اول و دوم، میزان شباهت بین این دو جمله محاسبه می‌شود. نحوه محاسبه شباهت کسینوسی بر اساس رابطه (۱) در ادامه توضیح داده شده است:

رابطه (۱):

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

در رابطه ۱ A و B بردارهای مربوط به جملات اول و دوم با n بعد می‌باشند.

نتایج

با استفاده از مدل سیامی و شبکه عصبیLSTM ، آموزش مدل با استفاده از داده‌های معرفی شده انجام شد. در این آزمایش از بردارهای با ابعاد ۱۰۰ از مدل araVec استفاده کردیم. داده‌ها پس از ترکیب به نسبت‌های ۸۰، ۱۰، ۱۰ تقسیم شدند: ۸۰ درصد برای آموزش، ۱۰ درصد برای اعتبارسنجی و ۱۰ درصد برای آزمون. تعداد نرون‌های مخفی در شبکه LSTM برابر با ۱۰۰ و تعداد گام‌های آموزشی ۵۰ بوده است. برای ارزیابی مدل، از معیار همبستگی پیرسون استفاده شد. این رابطه یکی از متداول‌ترین معیارها در آمار برای سنجش ارتباط بین مجموعه آزمون و داده‌های واقعی است. ضریب همبستگی پیرسون بین -۱ و ۱ تغییر می‌کند و در این پژوهش ضریب به ۵ ضرب شده است. طبق رابطه ۲، اگر r = 1 باشد، نشان‌دهنده رابطه مستقیم کامل بین دو متغیر است؛ بدین معنا که با افزایش (کاهش) یکی از متغیرها، دیگری نیز به‌طور مشابه افزایش (کاهش) می‌یابد. اگر- r = 1باشد، وجود رابطه معکوس بین دو متغیر را نشان می‌دهد؛ به این صورت که با افزایش یکی از

متغیرها، دیگری کاهش می‌یابد و بالعکس. زمانی که ضریب همبستگی برابر با صفر باشد، به این معنی است که بین دو متغیر رابطه خطی وجود ندارد.

رابطه (۲):

$$r = \frac{n\,(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

در این رابطه، n تعداد نمونه‌های مجموعه آزمون است. xمیزان شباهت تخمین زده شده و y میزان شباهت واقعی است که از داده‌های اصلی به‌دست می‌آید. نتایج بر اساس میزان همبستگی پیرسون در جدول ۲ آورده شده است. همان‌طور که در جدول ۳ مشاهده می‌کنید، استفاده از مدل CBOW در بازنمایی برداری کلمات عربی نتایج بهتری نسبت به مدل Skip-gram داشته است.

جدول (۳): نتایج بر اساس همبستگی پیرسون

روش	میزان همبستگی پیرسون
سیستم پیشنهادی CBOW vector	۸۳٫۴
سیستم پیشنهادیSkip-gram vector	۸۱٫۷
(وو و BIT با نام semEval2017 بهترین سیستم[1] همکاران، ۲۰۱۷)	۷۵٫۴۳
اسچواب و همکاران ۲۰۱۷	۷۹٫۶۹
بیجروا و همکاران، ۲۰۱۷	۷۱
تیان و همکاران، ۲۰۱۷	۷۴٫۴
بریچین، ۲۰۲۰	۶۹

همان‌طور که در جدول شماره (۲) مشاهده می‌کنید، استفاده از شبکه‌های LSTM در مدل Siamese Networkدر مقایسه با سایر روش‌ها از نقاط قوت بیشتری برخوردار است و خروجی برداری آن می‌تواند معنای کامل‌تری از جمله را نمایش دهد. پیکره‌های مورد استفاده در جدول (۲) مربوط به دادگان SEM2017 است. همچنین، با استفاده از دادگان بین‌زبانی عربی-انگلیسی، مدل شباهت‌یابی معنایی بین‌زبانی آموزش داده شد. در این روش، آموزش با استفاده از مدل مبتنی بر مبدل با نامparaphrase-xlm-r-multilingual صورت گرفت و نتیجه ۸۲ درصد به‌دست آمد.

1. arxiv.org/pdf/1708.00055.pdf.

نتیجه‌گیری

شباهت‌یابی معنایی متون یکی از زیرشاخه‌های پردازش زبان طبیعی است که در چند سال اخیر توجه زیادی را به خود جلب کرده است. در این مقاله، سیستمی برای محاسبه میزان شباهت معنایی بین دو جمله در زبان عربی با استفاده از یادگیری عمیق معرفی شده است. با توجه به اینکه بسیاری از متون اسلامی به زبان عربی هستند، این پژوهش از کاربردهای فراوانی برخوردار است. برای محاسبه میزان شباهت معنایی بین دو جمله، از بردارهای از پیش آموزش داده شده بر روی متون عربی موجود در توییتر استفاده شده است. بردارهای کلمات با استفاده از دو روش معروف CBOW و Skip-Gram استخراج می‌شوند که از روش‌های برجسته آموزش تعبیه کلمات هستند. پس از تولید بردارهای کلمات مربوط به جملات، با استفاده از لایه LSTM بردار جملات ایجاد شده و سپس با استفاده از میزان شباهت کسینوسی بین دو بردار، میزان شباهت جملات محاسبه می‌شود. استفاده از LSTM توانایی یادگیری وابستگی‌های بلندمدت را در شبکه فراهم می‌سازد. نتایج این تحقیق نشان می‌دهد که با استفاده از روش پیشنهادی، میزان همبستگی ۸۳/۴ درصد به‌دست آمده که عملکرد بهتری نسبت به سایر روش‌های موجود دارد. همچنین، با استفاده از دادگان بین‌زبانی عربی-انگلیسی ارائه شده در کنفرانس Semeva، مدل آموزش داده شد که میزان همبستگی ۸۲ درصد را به‌دست آورد.

استفاده از مدل‌های جدیدتر شبکه عصبی مانند Elmo، OpenAI-GPT و BERT به دلیل توجه عمیق‌تر به معنای جمله می‌تواند نتایج بهتری نسبت به روش پیشنهادی ارائه دهد. هرچه بازنمایی‌های تولید شده از جملات کیفیت بالاتری داشته باشد، عملکرد سامانه‌های شباهت‌یابی معنایی نیز بهبود می‌یابد. همچنین، اگر جملاتی که از لحاظ معنایی مشابه‌تر هستند در فضای برداری به یکدیگر نزدیک‌تر قرار گیرند، کیفیت سامانه نیز بهبود خواهد یافت

فهرست منابع

Agirre, Eneko, et al. (2012), "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).

Agirre, Eneko, et al. (2014), "Semeval-2014 Task 10: Multilingual Semantic Textual Similarity." Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014).

Agirre, Eneko, et al. (2013), "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.

Bar, D, Biemann, C, Gurevych, I, and Zesch, T. (2012), "UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures." Proceedings of the 6th International Workshop on Semantic Evaluation, in conjunction with the 1st Joint Conference on Lexical and Computational Semantics.

Bjerva, Johannes, and Robert Östling. (2017), "Cross-lingual Learning of Semantic Textual Similarity with Multilingual Word Representations." 21st Nordic Conference on Computational Linguistics, NoDaLiDa, Gothenburg, Sweden, 22-24 May. Linköping University Electronic Press.

Brychcín, Tomáš. (2020), "Linear Transformations for Cross-lingual Semantic Textual Similarity." Knowledge-Based Systems 187: 104819.

Cer, Daniel, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. (2017), "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation." Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics.

Comelles, Elisabet, and Jordi Atserias. (2019), "VERTa: A Linguistic Approach to Automatic Machine Translation Evaluation." Language Resources and Evaluation 53.1: 57-86.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. (2005), "The PASCAL Recognizing Textual Entailment Challenge." Machine Learning Challenges Workshop. Springer, Berlin, Heidelberg.

Das, Arijit, and Diganta Saha. (2022), "Deep Learning Based Bengali Question Answering System Using Semantic Textual Similarity." Multimedia Tools and Applications: 1-25.

Han, Lushan, et al. (2013), "UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems." Second Joint Conference on Lexical and Computational Semantics (Sem), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity.

Hochreiter, Sepp, and Jürgen Schmidhuber. (1997), "Long Short-Term Memory." Neural Computation 9.8: 1735-1780.

Islam, Aminul, and Diana Inkpen. (2008), "Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity." ACM Transactions on Knowledge Discovery from Data (TKDD) 2.2: 1-25.

Lubani, Mohamed, and Shahrul Azman Mohd Noah. (2019), "Text Relation Extraction Using Sentence-Relation Semantic Similarity." In Multi-disciplinary Trends in Artificial Intelligence: 13th International Conference, MIWAI 2019, Kuala Lumpur, Malaysia, November 17–19, Proceedings 13, pp. 3-14. Springer International Publishing.

Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. (2006), "Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity." AAAI Vol. 6. No. 2006.

Mikolov, Tomas, et al. (2013), "Distributed Representations of Words and Phrases and Their Compositionality." Advances in Neural Information Processing Systems.

Mueller, J, & Thyagarajan, A. (2016), "Siamese Recurrent Architectures for Learning Sentence Similarity." Thirtieth AAAI Conference on Artificial Intelligence.

Roul, Rajendra Kumar, and Jajati Keshari Sahoo. (2020), "Near-Duplicate Document Detection Using Semantic-Based Similarity Measure: A Novel Approach." In Computational Intelligence in Data Mining:

Proceedings of the International Conference on ICCIDM 2018, pp. 543-558. Springer Singapore.

Rychalska, B, Pakulska, K, Chodorowska, K, Walczak, W, and Andruszkiewicz, P,(۲۰۱۶) .

"Samsung Poland NLP Team at SemEval-2016 Task 1: Necessity for Diversity; Combining Recursive Autoencoders, Wordnet and Ensemble Methods to Measure Semantic Similarity." Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, CA, USA.

Schwab, Didier. (2017), "Semantic Similarity of Arabic Sentences with Word Embeddings".

Shahmirzadi, Omid, Adam Lugowski, and Kenneth Younge. (2019), "Text Similarity in Vector Space Models: A Comparative Study." 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE.

Sultan, M.A, Bethard, S, and Sumner, T. (2014), "Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence." Transactions of the Association for Computational Linguistics, 2:219–230.

Sultan, M.A, Bethard, S, and Sumner, T. (2014), "DLS@CU: Sentence Similarity from Word Alignment." Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 241–246, Dublin, Ireland, August.

Soliman, Abu Bakr, Kareem Eissa, and Samhaa R. El-Beltagy. (2017), "Aravec: A Set of Arabic Word Embedding Models for Use in Arabic NLP." Procedia Computer Science 117: 256-265.

Suleiman, Dima, Arafat Awajan, and Nailah Al-Madi. (2017), "Deep Learning-Based Technique for Plagiarism Detection in Arabic Texts." 2017 International Conference on New Trends in Computing Sciences (ICTCS). IEEE.

Tian, Junfeng, et al. (2017), "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).

Wu, Hao, et al. (2017), "BIT at SemEval-2017 Task 1: Using Semantic Information Space to Evaluate Semantic Textual Similarity." Proceedings of the 11th International Workshop on Semantic Evaluation.