تشابه‌یابی معنایی بین آیات قرآن با استفاده از معماری شبکه‌ی بازگشتی siamese

محمد جواد سعیدی زاده
دانشجوی کارشناسی ارشد هوش مصنوعی دانشگاه علم و صنعت ایران
m_saeedizade@comp.iust.ac.ir

علی سرآبادانی
دانشجوی کارشناسی ارشد مهندسی کامپیوتر قرآن کاوی رایانشی دانشگاه شهید بهشتی تهران
al.sarabadani@mail.sbu.ac.ir

نجمه ترابیان
دانشجوی دکتری هوش مصنوعی دانشگاه علم و صنعت ایران
Najmeh.torabian@gmail.com

چکیده

تشابه‌یابی بین جملات یکی از وظایف پردازش زبان طبیعی بوده تا ماشین، معنای جملات را درک کند. تشابه‌یابی بیان می‌کند دو جمله چقدر از لحاظ لفظ و معنا به یکدیگر شبیه هستند، با اینکار می‌توان موضوعات مرتبط با موضوع خود را به راحتی استخراج کرد. این کار با امتیاز دهی عددی بین ۰ تا ۵ است که (عدد ۰ بیانگر کاملا متفاوت و عدد ۵ بیانگر شباهت زیاد) میزان شباهت دو را بیان می‌کند. در این کار ما از دادگان انگلیسی مربوط به تشابه‌یابی جملات استفاده کرده و استفاده از دادگان قرآن را برای کارهای بعدی می‌گذاریم. می‌خواهیم مدلی که می‌سازیم قادر باشد دو جمله انگلیسی را بگیرد و میزان شباهت آن‌ها را یاد گرفته و بیان کند سپس مدل ساخته شده را برای پیدا کردن شباهت میان ترجمه انگلیسی آیات قرآن استفاده می‌کنیم. در این کار از معماری Siamese Recurren برای این شباهت یابی استفاده شده است.

کلمات کلیدی: پردازش زبان طبیعی، شباهت‌یابی جملات، قرآن‌کاوی رایانشی، یادگیری عمیق، قرآن

مقدمه

فهم متن یک وظیفه مهم در هوش مصنوعی برای درک معنای متن و پیدا کردن تشابه معنایی بین دو جمله یا متن می‌باشد. مدلی را مناسب می‌دانیم که بتواند معنای جملات در زمینه‌های متفاوت را بهتر درک کند. در این کار می‌خواهیم مدلی بسازیم که تشابه معنایی آیات قرآن را بیان کند. در مدل‌های یادگیری ماشین جمع آوری داده آموزشی کاری سخت و پر هزینه است از این رو ما روی دادگان موجود عملیات یادگیری و ساخت مدل را انجام میدهیم. دادگان موجود به زبان انگلیسی هستند و مدلی که ما ایجاد می‌کنیم قادر به یادگیری تشابه معنایی بین جملات انگلیسی خواهد بود از این رو ما ترجمه قرآن به زبان انگلیسی را برای ارزیابی کار خود استفاده میکنیم. در قدم اول ما یک قالب برای نمایش از ورودی‌های خود باید ایجاد کنیم که بتواند جمله اول و دوم را در قالب خود بیان کنند. ما ابتدا هر جمله را به لیستی از کلمات تبدیل می‌کنیم و سپس به دلیل اینکه شبکه یادگیری تنها اعداد را درک میکند هر کلمه را با استفاده از Google 300-dimensiona W2V (Mikolov, Tomas, et al.) به نمایش عددی تبدیل میکنیم. برای بیان جمله، ما لیستی از بردارهایی که نماینده کلمات هستند را داریم که دارای طول‌های متفاوتی می‌باشند و برای یکسان سازی طول تمامی جملات ما طول هر جمله را برابر بیشترین طول جمله کرده و با استفاده از zero padding پایان جملات را پر کرده‌ایم. از بین مدل‌های شبکه عصبی‌ای، LSTM (Gers, Felix A.) میتواند سری ها را یاد بگیرد زیرا میتواند روابط مهم و ناشناخته ای را بین اطلاعات داخل سری‌ها رو استخراج کند. حل این مسئله را یک مسئله یادگیری با ناظر در نظر گرفته و داده‌های آموزشی را که شامل جفت جملات می‌باشد به شکل $(x_1^a \ldots . x_{T_a}^a).(x_1^b \ldots . x_{T_b}^b)$ و برچسب Y که بیانگر میزان شباهت است و الگوریتم ما سعی در یادگیری این عدد دارد. طول جملات مربوط به a یا b ثابت بوده اما طول هر دو جمله a و b میتواند متفاوت باشد ($T_a \neq T_b$). در اینجا هر x_i بیانگر بردار هر کلمه بوده و شبکه LSTM (Gers, Felix A.) تلاش در استخراج این تشابه معنایی می‌کند.

کارهای مرتبط

در مورد شباهت معنایی در قرآن کریم در سطح یک آیه و حتی سوره‌های قرآن مقالاتی در گذشته کار شده است، به طور اجمالی اهداف و ایده‌های این مقالات را بررسی می‌کنیم:

در مقاله‌ای که متعلق به احسان خندگی (Khadangi, Ehsan, Mohammad Moein Fazeli) است در مورد یکپارچگی موضوعاتی سوره‌های قرآن بحث می‌کنند. ایده مطرح شده اینگونه است که مدعی است که عناصر درونی سوره‌ها رابطه تنگاتنگی با یکدیگر دارند و اینکه هر سوره از قرآن در یک موضوع اصلی ساخته شده است و قصد دارد تا یکسان بودن موضوع را در سوره‌های قرآنی با استفاده از زبان طبیعی بررسی کند. روشهای پردازش در این مقاله، براساس دو روش همراهی word2vec و Roots در آیات، شباهت ریشه‌های قرآنی محاسبه می‌کند.

در مقاله‌ای دیگر کاری از آقای Sameer M. Alrehaili (Alrehaili, Sameer M., Mohammad Alqahtani) می‌باشد در مورد ارائه روشی برای تراز کردن منابع معنایی قرآنی عربی هم بحث شده است.

در مورد ارزیابی تشابه‌های ترجمه‌های انگلیسی قرآن نیز آقای Mohd Zamri Murah (Murah, Mohd Zamri.) در مقاله اینگونه می‌گوید :

روش‌های محاسباتی برای ارزیابی عبارت از کلمات، اصطلاحات فرکانس، TF-IDF و نمایه سازی معنایی نهفته است. با استفاده از بیست و یک جفت ترجمه از هفت ترجمه انگلیسی Hilali ، YusufAli، Sahih، Shakir، Arberry، Pickthall، Maududi در شباهت به صورت دو طرفه مورد بررسی قرار گرفت. بر اساس نتایج‌هایشان، هفت جفت ترجمه شباهت بالایی دارند. (هیلال، یوسف عالی)، (هیلالی، صحیح)، (صحیح، شکیر)، (هیلالی، پی پتال)، (پیکتال، شکیر)، (هلالی، شکیر)، (شکیر، آربری).

برای یافتن مترادف کلمات قرآنی(عربی) نیز مقاله‌ای از Manal AlMaayah (AlMaayah, Manal) منتشر شده است، در این مقاله، یک مدل استخراج خودکار مترادف را تهیه کرده‌اند، که برای ساختن WordNet (Fellbaum, Christiane.) عربی قرآنی، به (QAWN) - فرهنگ لغات سنتی عربی- بستگی دارد. هدف از این کار پیوند دادن کلمات قرآنی معانی مشابه به منظور تولید مجموعه‌های مترادف (همگام سازی) است. برای تحقق این امر، از فرکانس اصطلاح و فرکانس اسناد معکوس در مدل فضای برداری استفاده کردند و سپس شباهت‌های بین کلمات قرآنی را بر اساس تعاریف متنی که از فرهنگ لغات سنتی عربی استخراج شده است محاسبه کردند. کلمات با بالاترین شباهت برای تشکیل یک شبکه مشترک با هم گروه بندی شدند.

در مقاله‌ای از El Moatez (Schwab, Didier.) در مورد یافتن تشابه معنایی جملات عربی به وسیله Word Embedding مطالبی مطرح کرده است.

معتقد است که شباهت متنی معنایی اساس برنامه‌های بی شماری است و در مناطق مختلف مانند بازیابی اطلاعات، کشف سرقت ادبی، استخراج اطلاعات و ترجمه ماشین نقش مهمی دارد. در این مقاله یک سیستم مبتنی بر devoted to calculate تعبیه شده که برای محاسبه شباهت معنایی در جملات عربی ارائه شده است. ایده اصلی بهره‌برداری از بردارها به عنوان بازنمایی کلمات در یک فضای چند بعدی به منظور گرفتن ویژگی‌های معنایی و نحوی کلمات است.

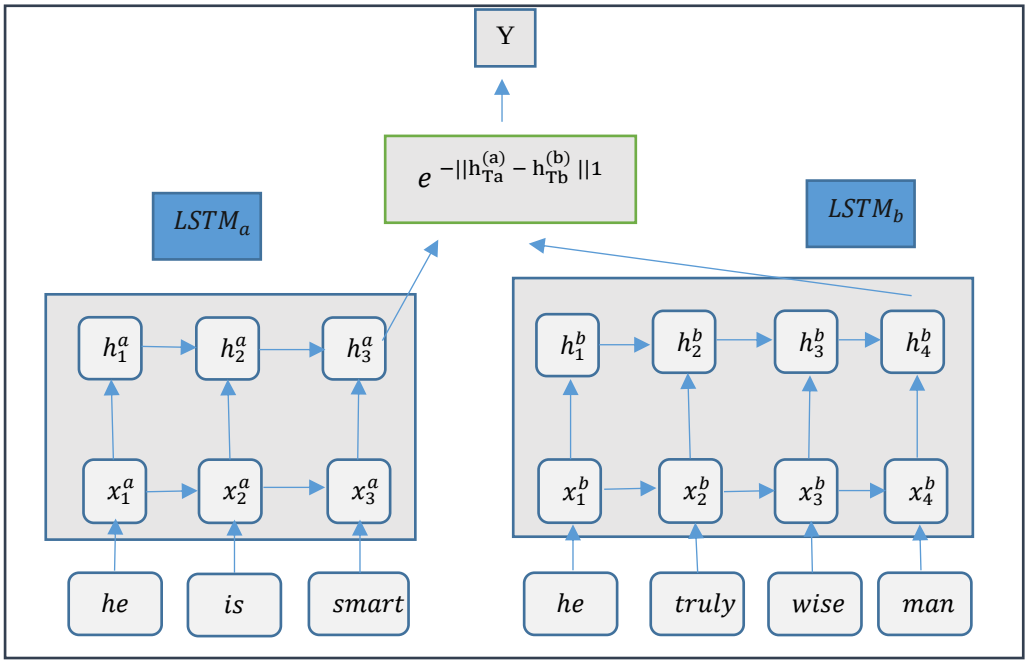
در مقاله ای دیگر از M. Akour (Akour, Mohammed, Izzat M. Alsmadi) در مورد اندازه گیری شباهت آیات قرآن و طبقه بندی سوره با استفاده از N-Gram مطالبی مطرح شده است که می‌گوید تلاش‌های گسترده تحقیقاتی در زمینه بازیابی اطلاعات در توسعه سیستم‌های بازیابی مربوط به زبان عربی برای روش‌های مختلف زبان طبیعی و بازیابی اطلاعات متمرکز شده است. با این حال، تلاش‌های اندکی در آن زمینه‌ها برای استخراج دانش از قرآن انجام شده است.

حتی کتاب‌هایی در این زمینه تالیف شده است که می‌توانیم به کتابی از آقای Salha Hassan Muhammed Qahl (Qahl, Salha Hassan Muhammed.) که در مورد موتور تشخیص تشابه خودکار بین متون مقدس با استفاده از روش کاوی و اندازه گیری شباهت آن‌ها، اشاره کنیم. ایشان در این کتاب تفاوت بین استخراج ویژگی‌های نحوی مبتنی بر دامنه و استخراج ویژگی‌های بدون دامنه را ارزیابی کردند و سپس از انواع اقدامات شباهت مانند Euclid، Hilling، Manhattan، Cassin Batcharia، kullback-leibler symmetries و Jensen Shannon احتمالی استفاده کردند و همچنین از chi-square and clark، برای شناسایی شباهت ها و تفاوت های بین متون مقدس استفاده کردند.

روش پیشنهادی

برای درک شباهت معنایی بین جملات از روش Manhattan LSTM (Mueller, Jonas) که در شکل ۱ نمایش داده شده استفاده میکنیم. LSTM (Gers, Felix A.) یک معماری شبکه عصبی بازگشتی مصنوعی است که در زمینه یادگیری عمیق مورد استفاده قرار می‌گیرد. برخلاف شبکه‌های عصبی استاندارد feed-forward، اتصالات بازگشتی دارد. این نه تنها می‌تواند نقاط داده‌های واحد را پردازش کند، بلکه توالی‌های کل داده را نیز پردازش می‌کند. در این شبکه از دو شبکه بازگشتی استفاده شده که $LSTM_a$ و $LSTM_b$ سعی در یادگیری شباهت معنایی را دارند. کار هر شبکه مپ کردن بردار نمایش جمله $(x_1 \ldots . x_T)$ که هر x_i بردار کلمه در فضای ۳۰۰ بعدی مدل W2V google (Mikolov, Tomas, et al.) به فضای برداری به طول ۵۰ می‌باشد. word2vec (Mikolov, Tomas, et al.) از یک روش معمول در فهم متون و یا لغات ناآشنا الهام گرفته شده است. یعنی این ایده که در هنگام مواجهه با یک لغت ناآشنا در متن (به‌خصوص متون زبان خارجی) با توجه به سیاق مطلب و سایر واژگان همسایگی لغت ناآشنا، می‌توان تا حدود بسیار خوبی تقریبی از معنی و مفهوم آن واژه و یا حتی نقش آن

به دست آورد. راهکار ارائه‌شده Word2Vec تبدیل واژگان به بردار و انتقال آن به فضای برداری است که امکان پردازش لغات و متن‌ها را با ابزارهای یادگیری ماشین امکان‌پذیر و آسان می‌سازد.



شکل ۱ : مدل ما از LSTM (Gers, Felix A.) برای خواندن بردارهای کلمات استفاده می‌کند که هر کلمه را به حالت پنهان نهایی خود معرفی می‌کند و آن را به عنوان یک بردار به کار می‌برد. بردار تعبیه شده برای هر جمله، شباهت بین این بازنمایی ها را به عنوان پیش بینی کننده شباهت معنایی استفاده می‌کند.

از شبکه انتظار داریم که اگر دو جمله‌ی تقریبا متشابه به عنوان ورودی بگیرد بردارهایی که در خروجی تولید می‌کند، به هم نزدیک باشند. قرینه فاصله ی منهتن این دو بردار را حساب کرده و به توان e رسانده است، و آن را به عنوان خروجی گزارش کرده است:

$$g(h_{Ta}^{(a)}, h_{Tb}^{(b)}) = e^{-||h_{Ta}^{(a)} - h_{Tb}^{(b)}||1} \in [0, 1]$$

برای آموزش شبکه ازمعیار خطای MSE به معنای میانگین مربع خطا و بهینه ساز ADAM (Kingma, Diederik P.) استفاده شده است.

پس از آموزش شبکه ما ترجمه انگلیسی آیات قرآن را به فرمت ورودی شبکه تبدیل و به شبکه تزریق میکنیم، خروجی شبکه میزان شباهت دو آیه می‌باشد که در بخش ارزیابی خروجی در عدد ۵ ضرب شده است. باید تمام جفت آیات ممکن به شبکه داده شود و قابل ذکر است که مرتبه زمانی این کار n^2 بوده، برای سادگی فقط جزء سی قرآن در این کار استفاده شده است.

ارزیابی

دادگان STS شامل ۸۶۲۸ جفت جمله که به سه بخش آموزش ۵۷۴۹، اعتبارسنجی ۱۵۰۰ و تست ۱۳۷۹ جمله تقسیم بندی شده است. به هر جفت جمله یک عدد بین صفر تا پنج به عنوان میزان شباهت نسبت داده شده است. در جدول ۱ برخی از نتایج مدل را روی داده تست مشاهده می‌کنید:

جمله دوم	جمله اول	تخمین مدل	برچسب داده
A young child is riding a horse.	A child is riding a horse.	4.1	4.75
A woman peels a potato.	A woman is peeling a potato.	4	5
Three men are playing guitars.	Three men are on stage playing guitars.	3.8	3.75
A woman is carrying a boy.	A woman is carrying her baby.	1.9	2.33
A man is erasing a chalk board.	The man is erasing the chalk board.	2.8	5

جدول ۱ : اجرای مدل siamese روی بخش تست دادگان sts. ستون برچسب داده بیانگر مقداری است که مدل می‌بایستی تخمین میزد.

در جدول ۱ می‌بینید که برای آموزش مدل از دادگان انگلیسی استفاده شده و تخمین مدل به مقدار واقعی آن در بیشتر می‌باشد.

بخش تعمیم مدل به آیات قرآن:

آیه دوم	آیه اول	تخمین تشابه بر اساس فاصله کسینوسی	تخمین شباهت مدل MALSTM
فَإِنّ مَعَ الْعُسْرِ يُسْرًا	إِنّ مَعَ الْعُسْرِ يُسْرًا	4.96	**4.7**
أَلَا يَظُنّ أُولَئِكَ أَنّهُمْ مَبْعُوثُونَ	وَكُلّ شَيْءٍ أَحْصَيْنَاهُ كِتَابًا	3.25	**3.77**
وَسَيُجَنّبُهَا الْأَتْقَى	وَيَتَجَنّبُهَا الْأَشْقَى	4.6	**4**
فَإِنّ الْجَحِيمَ هِيَ الْمَأْوَى	إِنّهُمْ كَانُوا لَا يَرْجُونَ حِسَابًا	3.45	**3.51**
عَلِمَتْ نَفْسٌ مَا أَحْضَرَتْ	كَلّا سَيَعْلَمُونَ	1.1	**3**

جدول ۲ : اجرای siamese (Mueller, Jonas) روی ترجمه انگلیسی قرآن و نمایش عربی آیات و شباهت آنها. در ستون تشابه بر اساس فاصله کسینوسی، بجای فاصله منهتن از ضرب کسینوسی برداری استفاده شده است.

در جدول ۲ می‌بینید پس از آموزش مدل روی دادگان انگلیسی روی ترجمه قرآن خیلی خوب تشابه معنایی آیات را یاد گرفته است.

کارهای بعدی

در این کار شباهت‌یابی آیات در جزء سی‌ام قرآن انجام شد و دلیل آن این بود که پیچیدگی الگوریتمی که دوبه‌دوی آیات را بررسی می‌کرد از مرتبه زمانی $O(n^2)$ بود. بررسی تمام آیات را برای کارهای بعدی میگذاریم. بررسی شباهت احادیث و آیات قرآن نیز می‌تواند در کارهایی که در راستای این مقاله انجام می‌شود قرار گیرد.

مراجع

[1] Mikolov, Tomas, et al. "**Distributed representations of words and phrases and their compositionality.**" Advances in neural information processing systems. 2013.

[2] Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "**Learning to forget: Continual prediction with LSTM.**" (1999): 850-855.

[3] Mueller, Jonas, and Aditya Thyagarajan. "**Siamese recurrent architectures for learning sentence similarity.**" thirtieth AAAI conference on artificial intelligence. 2016.

[4] Kingma, Diederik P., and Jimmy Ba. "**Adam: A method for stochastic optimization**." arXiv preprint arXiv:1412.6980 (2014).

[5] Khadangi, Ehsan, Mohammad Moein Fazeli, and Amin Shahmohammadi. "**The Study on Quranic Surahs' Topic Sameness Using NLP Techniques.**" 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE, 2018.

[6] Alrehaili, Sameer M., Mohammad Alqahtani, and Eric Atwell. "**A Hybrid Methods of Aligning Arabic Qur'anic Semantic Resources**." 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR). IEEE, 2018.

[7] Murah, Mohd Zamri. "**Similarity Evaluation of English Translations of the Holy Quran.**" 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences. IEEE, 2013.

[8] AlMaayah, Manal, Majdi Sawalha, and Mohammad AM Abushariah. "**Towards an automatic extraction of synonyms for Quranic Arabic WordNet.**" International Journal of Speech Technology 19.2 (2016): 177-189.

[9] Fellbaum, Christiane. "**WordNet**." The encyclopedia of applied linguistics (2012).

[10] Schwab, Didier. "**Semantic similarity of arabic sentences with word embeddings.**" 2017.

[11] Akour, Mohammed, Izzat M. Alsmadi, and Iyad Alazzam. "**MQVC: Measuring quranic verses similarity and sura classification using N-gram**." (2014).

[12] Qahl, Salha Hassan Muhammed. "**An Automatic Similarity Detection Engine Between Sacred Texts Using Text Mining and Similarity Measures**." (2014).