Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Sentence Embedding Using Transformer Encoder for Retrieving Answers with Higher Accuracy to User Queries

Godavarthi Deepthi [1][0000-0003-0712-6899], **A. Mary Sowjanya** [2][0000-0003-2217-3925]
[1] Assistant Professor, School of Computer Science and Engineering, VIT-AP University-Amaravati, Andhra Pradesh, India.
[2] Associate Professor, Department of Computer Science and Systems Engineering, Andhra University College of Engineering(A), Visakhapatnam, Andhra Pradesh, India

Article Info
Page Number: 2424 – 2429
Publication Issue:
Vol 71 No. 4 (2022)

Article History
***Article Received**: 25 March 2022*
***Revised**: 30 April 2022*
***Accepted**: 15 June 2022*
***Publication**: 19 August 2022*

Abstract
Word embeddings are used for several Natural Language Processing (NLP) tasks but these are not effective for obtaining embeddings for sentences. Sentence embeddings can be used to resolve this issue. They provides the vector representation for sentences along with semantic information from which the machine gets clear understanding about the context. In this work, Question Answering system using universal sentence encoder (USE) with transformer encoder variant USE_{Trans} is developed to extract the sentence with the correct answer for a user query from the context. If the sentence having correct answer is identified efficiently it would be much more easier to retrieve the exact answer from the sentence.The developed model helps to provide exact answer to user query. The developed model is evaluated on SQuAD-2.0 dataset. Compared to $USE_{DAN,}$ it is observed that USE_{Trans} yields better accuracy.

Introduction :

In word embedding, words are represented using vectors. It is very hard to obtain information from word embeddings if the text size is very large. These are not successful in generating embeddings for large text segments like sentences. For instance in a sentence such as 'I don't like cramped locations' and another sentence 'Even though Exhibition is overflowing with people I like it, becomes difficult for a Machine to find out the variation between 'cramped' and 'overflowing'. Sentence embeddings are effective to resolve these issues. Full sentences and the semantic information is denoted as vectors through which machine gets a better understanding on the context.In this work, Transformer encoder variant(USETrans) variant of universal sentence encoder is used to build a question answering system that extracts the sentence having the relevant answer for a query from the paragraphs present in the dataset. Stanford Question Answering Dataset 2.0 (SQuAD-2.0) has been used in this work. It has queries on articles uploaded in Wikipedia by various several workers. Response to every query is a small text segment from the context.

Literature Survey:

Buzaaba et al.[1] broken the QA problem into three components namely entity detection, linking and relation prediction components. Neural network is used for entity detection ,non-neural network methods are used for relation prediction and some heuristics are used for entity linking.

Lucy Lu Wang and Lo [2] described various resources introduced to provide support to text mining applications on COVID research work. They discussed the corpora, resources, shared tasks and systems were introduced for COVID-19. A total of 39 systems that contributed functionalities like text visualization and text summarization on COVID-19 research work are compiled by them.
Yang et al.[3] presented multilingual sentence embedding models that used multi task trained dual encoder. This is used to embed the data from several languages into shared semantic space. Multilingual embeddings performed better than English only sentence embeddings in some cases.
J. Lee et al. [4] created a question answering framework, COVIDASK, that concatenated biomedical text mining along with Question Answering(QA) methods to produce responses to real-time queries.
Barros et al. [5] developed a novel semantic-based pipeline that recommends biomedical entities to research community. Based on Named Entity Recognition, the designed pipeline used multidisciplinary ontologies to generated a feedback matrix using for recognition and link the entities.
A. Amini et al. [6] studied several techniques to retrieve various forms of mechanism relations from scientific related papers. They introduced a coarse-grained schema selection technique relations among open and free-form entities.
E. Ogundepo et al.[7] introduced a dataset on COVID-19 updates which was announced online by the Nigeria Centre for Disease Control (NCDC) from 27th february to 29th september in 2020. Web scraping was done from several sources to obtain the data.
Arantxa Otegi et al. [8] developed a Question Answering (QA) framework ,an integration of an Information Retrieval component with reading comprehension component that produce responses from the retrieved paragraphs.
Kirk Roberts et al[9] developed an information retrieval task, TREC-COVID for providing support to clinicians and clinical research during the pandemic. It is different from other information retrieval shared tasks with remarkable considerations.
Manivannan [10] studied different closed domain question answering(QA) approaches. The developed question answering system on Hyderabad tourism provide answers to questions about city history, monuments, parks, lakes of Hyderabad city.
Yu Hao et al.[11] developed supervised learning architecture using sentence embeddings for question answering system on medical domain. The sentence embedding producing module is used to measure similarity while scoring module to capture association among sentence pairs.

Methodology :

The two variants of universal sentence encoder have been used in this work to extract answer for a given user query from the context. SQUAD-2.0 dataset[12] is considered to be an open domain dataset as the questions in it are not restricted to a particular domain but contains questions from multiple domains.The task is to identify the sentence having the correct answer.

Universal sentence encoder

Usually for tasks like text classification, semantic similarity etc text needs to be encoded into high dimensional vectors that can be done using Universal sentence encoder. There are two variants in universal sentence encoder namely Transformer encoder and Deep Averaging Network. This encoder on providing variable length text as input produces vector of 512 dimensional as output. Transformer enocoder is used to train the model.First install Tensorflow and Tensorflow hub to use universal sentence encoder.Load the model from TFhub. Take the reading comprehension from the dataset as shown in figure 1 and use TextBlob to break it into various sentences as shown in figure 2.

Question: Who managed the Destiny's Child group?

['Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ bee-YON-say) (born September 4, 1981) is an American singer, songwriter, record producer and actress.', "Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child.", **"Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time."**, 'Their hiatus saw the release of Beyoncé\'s debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".']

Figure 1: Comprehension from dataset

['Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ bee-YON-say) (born September 4, 1981) is an American singer, songwriter, record producer and actress.',
"Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child.",
"Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time.",
'Their hiatus saw the release of Beyoncé\'s debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".']

Figure 2: Comprehension broken into sentences

Calculate cosine similarity for sentence- query pair to create features after obtaining vector representation for all sentences which is shown in Figure 3 .

Context	question	id	answer start	text	sentences	target	sent_emb	quest_emb	Cosine_sim
Beyoncé Giselle Knowles-Carter (/bi:'jonseI/b..	When did Beyoncé start becoming popular?	56be85543aeaaa14008c9063	269	In the late 1990s	[Beyoncé Giselle Knowles-Carter (/bi:'jonseI/b..	1	[[0.0914527 3, 0.19478364, 0.156293464, …	[[0.15324688, 0.072615382, 0.154161462, …	[0.6543244122675963 8, 0.4489008188247680 7,…
Beyoncé Giselle Knowles-Carter (/bi:'jonseI/b..	What areas did Beyoncé compete in when she was…	56be85543aeaaa14008c9065	207	singing and dancing	[Beyoncé Giselle Knowles-Carter (/bi:'jonseI/b..	1	[[0.09145273, 0.19478364, 0.156293464, …	[[0.145826 32, 0.14263425, 0.07245158,…	[0.57365662524724231, 0.52272474674284492, …
Beyoncé Giselle Knowles-Carter (/bi:'jonseI/b..	When did Beyoncé leave Destiny's Child and bec…	56be85543aeaaa14008c9066	526	2003	[Beyoncé Giselle Knowles-Carter (/bi:'jonseI/b..	3	[[0.09145273, 0.19478364, 0.156293464, …	[[0.06152316, 0.15842514, 0.06432766,…	[0.4962272542451327, 0.46174627742637813, …

Figure 3: Generated Sentence embeddings, question embeddings and cosine similarity values

After loading the model, provide the sentence as input to it from which 512 dimensional vector is obtained as output. Sentence similarity is obtained between the sentences using the embeddings generated earlier. Obtain the target labels by transforming the text to index of the sentence having the text. As the dataset has most of the paragraphs with 10 or below 10 sentences, paragraph length is restricted to 10 sentences .Hence 10 labels are considered for prediction. A single feature is built for every sentence using cosine similarity. A few paragraphs have sentences below 10 for which feature value is replaced with 1 as the maximum cosine value is 1 is shown in Figure 4 .

Column _cos_0	Column _cos_1	Column _cos_2	Column _cos_3	Column _cos_4	Column _cos_5	Column _cos_6	Column _cos_7	Column _cos_8	Column _cos_9	target
0.576431	0.549713	0.556853	0.551321	1.0	1.0	1.0	1.0	1.0	1.0	1
0.496572	0.434304	0.462341	0.455274	1.0	1.0	1.0	1.0	1.0	1.0	1
0.399315	0.371276	0.446209	0.394241	1.0	1.0	1.0	1.0	1.0	1.0	3

Figure 4. Fill missing values with 1

Compare the generated sentence embeddings with the question embedding and the sentence having shortest distance from the query is identified using cosine similarity. Target labels are generated using cosine similarity scores for all the questions. Match the target label generated from the dataset with the lables generated using cosine similarity. This technique yielded an accuracy of 72%.

Dependency Parsing : Dependency parse tree is used as another essential feature in this work to increase the performance of the developed model. For navigation through the tree Spacy tree parsing is used. Constructed Parse tree for question and sentences are shown in Figure 5.

Question:
Who managed the Destiny's Child group?

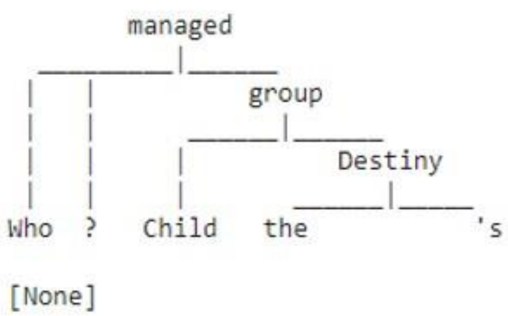

Figure 5. Generated Parse tree for question

Sentence having answer:

"Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time.", The answer is shown in Figure 6.

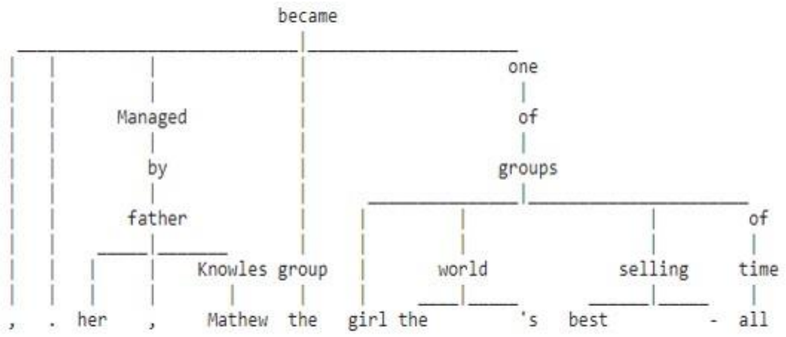

Figure 6. Generated parse tree for sentence having answer

The query root is matched with roots or subroots of a sentence.As there are more number of verbs in the sentence we will get several roots.There is high chance for finding an answer to the question from a sentence if query root exists in the root/subroot of a sentence.The feature corresponding to a sentence can be either 0 or 1.The feature is denoted with 1 if the root of the question exists in sentence roots else 0.Perform stemming before comparing question root with sentences roots. Various machine learning algorithms like XGBoost,Multi layer perceptron(MLP), Support Vector Machine(SVM),Logistic Regression,K-nearest neighbors(KNN),Random forest are used for training. The developed model gives good results on these algorithms compared to deep averaging network(DAN) which is shown in Table 1.

Table 1. Obtained accuracies with the two variants of universal sentence encoder

Machine Learning Algorithms	Universal sentence encoder with Deep Averaging Network (USE$_{\text{DAN}}$)	Universal sentence encoder with Transformer Encoder (USE$_{\text{Trans}}$)
XGBoost	**69.9**	**81.5**
MLP Classifier	**69.7**	**81.1**
SupportVector Machine	69.5	80.7
Logistic Regression	69.5	80.2
RandomForest classifier	69	80.2
K-nearest neighbor	61.2	75.6

Conclusion:

Answers for user queries can be retrieved from a context more efficiently using sentence embedding rather than word embedding. In this work, Transformer encoder (USE$_{\text{Trans}}$) a variant of universal sentence encoder has been used to develop a question answering system for extracting the sentence with correct answer. Sentence extraction proved to be better to retrieve the exact answer for a given question. The developed system has been found to perform well compared to USE_{DAN}, another variant of universal sentence encoder, when combined with XGBoost, Multilayer perceptron (MLP), Support Vector Machine (SVM), Logistic Regression, K-nearest neighbors (KNN), Random forest.

References:

[1] H. Buzaaba and T. Amagasa, "Question Answering Over Knowledge Base: A Scheme for Integrating Subject and the Identified Relation to Answer Simple Questions," *SN Comput. Sci.*, vol. 2, no. 1, pp. 1–13, 2021, doi: 10.1007/s42979-020-00421-7.

[2] L. L. Wang and K. Lo, "Text mining approaches for dealing with the rapidly expanding literature on COVID-19," *Brief. Bioinform.*, vol. 22, no. 2, pp. 781–799, 2021, doi: 10.1093/bib/bbaa296.

[3] Y. Yang *et al.*, "Multilingual Universal Sentence Encoder for Semantic Retrieval," pp. 87–94, 2020, doi: 10.18653/v1/2020.acl-demos.12.

[4] J. Lee *et al.*, "Answering Questions on COVID-19 in Real-Time," *arXiv*, 2020, doi: 10.18653/v1/2020.nlpcovid19-2.1.

[5] M. A. Barros, A. Lamurias, D. Sousa, P. Ruas, and F. M. Couto, "COVID-19: A Semantic-Based Pipeline for Recommending Biomedical Entities," 2020, doi: 10.18653/v1/2020.nlpcovid19-2.20.

[6] A. Amini *et al.*, "Extracting a Knowledge Base of Mechanisms from COVID-19 Papers," pp. 1–9, 2020.

[7] E. Ogundepo *et al.*, "An exploratory assessment of a multidimensional healthcare and economic data on COVID-19 in Nigeria," *Data Br.*, vol. 33, 2020, doi: 10.1016/j.dib.2020.106424.

[8] A. Otegi, J. A. Campos, G. Azkune, A. Soroa, and E. Agirre, "Automatic Evaluation vs. User Preference in Neural Textual QuestionAnswering over COVID-19 Scientific Literature," 2020, doi: 10.18653/v1/2020.nlpcovid19-2.15.

[9] K. Roberts *et al.*, "TREC-COVID: Rationale and structure of an information retrieval shared task for COVID-19," *J. Am. Med. Informatics Assoc.*, vol. 27, no. 9, pp. 1431–1436, 2020, doi: 10.1093/jamia/ocaa091.

[10] S. B. R. Manivannan, "A study on different closed domain question answering approaches," *Int. J. Speech Technol.*, vol. 23, no. 2, pp. 315–325, 2020, doi: 10.1007/s10772-020-09692-0.

[11] Y. Hao, X. Liu, J. Wu, and P. Lv, "Exploiting sentence embedding for medical question answering," *33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019*, pp. 938–945, 2019, doi: 10.1609/aaai.v33i01.3301938.

[12] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, pp. 784–789, 2018, doi: 10.18653/v1/p18-2124.