Training Doc2Vec on a Corpus of Persian Poems to Answer Thematic Similarity Multiple-Choice Questions

Soroosh Akef
Languages and Linguistics Center
Sharif University of Technology
Tehran, Iran
sor.akef@student.sharif.ir

Mohammad Hadi Bokaei
Department of Information Technology
Iran Telecommunication Research Center
Tehran, Iran
mh.bokaei@itrc.ac.ir

Hossein Sameti
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
sameti@sharif.edu

2020 10th International Symposium on Telecommunications (IST) | 978-1-7281-8012-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/IST50524.2020.9345918

***Abstract*—This paper reports our improvement over the previous benchmark of the task of answering poetic verses' thematic similarity multiple-choice questions (MCQs). These questions, which frequently appear in the Iranian national university entrance exam, require the test-taker to select the verse which is thematically similar to the stem verse. These questions would test a natural language processing (NLP) model's ability to not only understand, but interpret a poetic verse. In this experiment, we trained a Doc2Vec model on a corpus of Persian poems and proceeded to use the trained model to get the vector representations of the poetic verses. Subsequently, the poetic verse among the options with the highest cosine similarity to the stem verse was selected as the correct answer by the model. This model managed to answer %38 of the questions correctly, which was an improvement of %6 over the previous benchmark. Furthermore, it was observed that the model's approach to answering the questions seemed different from the previous model, as there were relatively few questions that the models had both answered correctly. The ultimate goal of this task is to estimate the difficulty of such questions by using the performance of the models as a feature. The fact that the two models did not behave similarly can prove useful, as the model whose performance has the most correlation with that of a human test-taker can be used for the future task of thematic similarity MCQ difficulty estimation.**

***Index Terms*—Doc2Vec, MCQ answering, natural language processing, poetry, question difficulty estimation**

I. Introduction

The last few years have seen rapid progress in the development of educational applications and websites. While significant work has been done with regard to utilizing artificial intelligence (AI) in education, taking advantage of AI to aid the process of educational material creation is rarely explored.

One of the most important matters in education is testing, as high-stakes tests can often shape the future of a student. Test difficulty, in particular, is one of the deciding factors in test validity and reliability. However, determining the difficulty of a test item is a laborsome and time-consuming task, which requires reliable piloting.

As a result, developing an intelligent system which estimates the difficulty of a question for an educator or test creator goes a long way in making future tests fairer. While numerous approaches may be adopted and a variety of features could be exploited in order to design an AI-driven system capable of determining the difficulty of a question, previous research has shown that there could be a correlation between the ability of an intelligent system and that of a student to answer a question [1]. It is for this reason that the current work has focused on designing an intelligent system which is capable of answering the type of questions whose difficulty we intend to measure.

Considering that the university entrance exam in Iran is the most high-stakes exam in this country, we have focused our attention on one type of question which constitutes a significant portion of the Persian Literature section of this exam. With approximately 9 of the 25 Persian Literature questions being poetic verses' thematic similarity multiple-choice questions (MCQs), these questions are popular among educational material creators [2].

Answering these questions, however, is not an easy task for an intelligent system, as it would ideally require a system to interpret poetic verses. That being said, it was shown that with the help of sentence embeddings generated using the pre-trained multilingual BERT[1] model, an intelligent system would be able to attain an accuracy of %32, which was a %7 improvement over the random guess baseline [3]. In the present work, we have attempted to improve on that performance by training a Doc2Vec model on a corpus of Persian poems.

In the subsequent sections, previous efforts concerning MCQ answering and Persian poems processing are discussed, the data used for the current experiment are described, and the experiment itself is further explained. In the final sections, the results obtained from the experiment and their implications are discussed, and some ideas for future research are presented.

978-1-7281-8012-0/20/$31.00 ©2020 IEEE

[1]Bidirectional Encoder Representations from Transformers

II. Related Work

A. MCQ Answering

The task of automatically answering MCQs has been receiving increasing attention in recent years. It has been argued that the ability to answer questions which require general knowledge about the world would be an indicator of the sophistication of an AI system, as current AI systems are often domain-specific [4].

The task of multiple-choice question answering has seen impressive results on reading comprehension questions with the best systems attempting the task of answering reading comprehension MCQs having attained above-human performance results. For instance, The Standford Question Answering Dataset (SQuAD2.0) requires a model to find the answer to a reading comprehension question in a text and abstain from answering when the answer is not found in the text. The human performance for this task has an F1 score of 89.452 [5], while the best current model has achieved an F1 score of 92.978 [6].

Such success, however, has been elusive in MCQ answering tasks which require general world knowledge. The Allen AI Science Challenge required an intelligent system to answer science questions typically given to an eighth-grader. The best models scored just under %60 and heavily depended on information retrieval (IR). [4]

More recently, attempts have been made to answer medical exam MCQs without the aid of any MCQ training data. An accuracy of %39.6 was obtained on a dataset of six-option MCQs. This experiment also took advantage of information retrieval to achieve the best result [7].

The current paper is an improvement over [3], which attempted to answer Persian poetic verses' thematic similarity MCQs simply with the help of embeddings obtained from the pre-trained multilingual BERT model. The accuracy of that model answering 100 four-option MCQs was %32, and the model displayed an inability to answer questions when the verses lacked semantic hints which the model could exploit.

B. Persian Poems Processing

Another aspect of the current task is applying natural language processing (NLP) techniques to Persian poems, which is an area that has room for further exploration. In this section, previous research conducted on Persian poems processing is discussed.

To the best of the authors' knowledge, the first work attempting to take advantage of natural language processing techniques to analyze Persian poems was [8]. This work attempted to cluster approximately 18,000 Persian ghazals[2] by 30 different poets using probabilistic topic modeling.

In another work, which also utilized probabilistic topic modeling, Hafez's[3] ghazals were classified chronologically using a support vector machine (SVM) classifier. [9] Subsequently, the features used in this work were expanded by introducing word embeddings and other innovative features in order to cluster Hafez's ghazals. [11]

[2]A type of Persian poem

[3]A prominent 14th century Persian poet

III. Data

As thematic similarity MCQs are among the most challenging questions for students, there are numerous supplementary materials available which contain MCQs for students to practice. This experiment uses the same dataset used in [3] in order to make a comparison between the results possible.

This dataset contains 100 thematic similarity MCQs stored in an Excel file with the first column containing the stem verse and the second to fifth columns containing option verses. The challenge of this type of MCQ is to find the option whose verse is thematically more similar to the stem verse.

In an actual exam, thematic similarity questions are of various types and may ask a test-taker to select the option which is different from the stem. However, in order to keep the dataset homogeneous, we have only included the so-called type-one questions, which require a test-taker to select the option most similar to the stem. A screenshot of the dataset is presented in Fig. 1.

It is worth noting that semantic similarity by itself is not enough to answer these questions, as many incorrect options contain similar words to the stem in order to distract the test-taker. The correct answer to each question is stored in a separate column.

The poetic verses used in the stem are often different from the verses used in the options, inasmuch as the verses used in the stem are usually selected from materials students are already familiar with while options' verses may be selected from unknown sources. Moreover, the stem may contain prose or Quranic verses at times while the options almost always contain a poetic verse.

The fact that the multilingual BERT model was trained on a corpus containing texts from Wikipedia means that the poetic language used in these MCQs is quite different from BERT's training corpus, and as a result, a model trained specifically on a corpus of Persian poetry would, in theory, yield better results.

IV. Method

In order to answer thematic similarity MCQs, we first trained a Doc2Vec model on a corpus of Persian poetry. Subsequently, the trained model was used to obtain vector representations for the verses of the stems and the options respectively. In the final stage, these vector representations were compared using cosine similarity in order to determine the correct answer.

Doc2Vec is an unsupervised algorithm which is used to vectorize documents in such a way that these vectors are representative of those documents. This algorithm is based on the earlier Word2Vec algorithm and attempts to encode the ordering of the words in a given sentence. After its release, Doc2Vec attained state-of-the-art results on numerous tasks [10].

Stem	Option 1	Option 2	Option 3	Option 4	Answer
گفت نزدیک است والی را سرای آن‌جا شویم گفت والی از کجا در خانه خمار نیست	در وجه معاش تو براق که نوشتند تغییر نیابد که ز دیوان الست است	ببینی مطلب زآن که درست است یقینم کآن خامه که این نقش نگارید شکسته‌ست	با محتسبم عیب مگویید که او نیز پیوسته چو ما در طلب عیش مدام است	آن کس که جویی و گلیمیش به دست است گر زین دو فزون می‌طلبد آزپرست است	3
گفت آگه نیستی کز سر درافتادت کلاه گفت در سر عقل باید پی‌کلاهی عار نیست	فکند از سر گردن‌کشان عالم خاک کلاه عقل تماشای طاق ابرویش	سری که در ره او پی‌کلاه می‌گردد فلک‌سوار چو خورشید و ماه می‌گردد	خرد باید اندر سر مرد و مغز نباید مرا چون تو دستار نغز	خورشید فلک را که جهان زیر نگین است جز خاک کف پای تو بر سر کلاهی نیست	3
بگفتا دل ز مهرش کی کنی پاک بگفت آن‌گه که باشم خفته در خاک	از مرگ نیندیشم گر جان به تو پیوندد پیری چه زیان دارد گر عشق جوان استی	آن را که زندگیش به عشق است مرگ نیست هرگز گمان مبر که مر او را فنا بود	در فراقت پی‌قرارم مرگ به زین زندگی جز غمت در دل ندارم مرگ به زین زندگی	چنان بربود خواب من که ناید چشم من بر هم مگر وقتی که زیر خاک خفته در کفن باشم	4
شور شراب عشق تو آن نفسم رود ز سر کاین سر پرهوس شود خاک در سرای تو	کسی که عشق نورزد سیاه‌دل باشد چو سر ز خاک لحد برزند خجل باشد	همه ذرات جهان مضطرب عشق تواند خاک را چون فلک از عشق تو آرامی نیست	نه من آنم که برگیرم سر از خاک درت هرگز مگر وقتی که زیر خاک خشتم زیر سر باشد	کسی کز سوز عشق تو ندارد جان و دل زنده به‌سان خاک گورستان درون پر مردگان دارد	3
بگفتا جان‌فروشی در ادب نیست بگفت از عشق‌بازان این عجب نیست	گر قلب دلم را ننهد دوست عیاری من نقد روان در دمش از دیده شمارم	پروانه او گر رسدم در طلب جان چون شمع همان دم به دمی جان بسپارم	حافظ لب لعلش چو مرا جان عزیز است عمری بود آن لحظه که جان را به لب آرم	امروز مکش سر ز وفای من و اندیش زآن شب که من از غم به دعا دست برآرم	2
بگفت آنجا به صنعت در چه کوشند بگفت انده خرند و جان فروشند	مه فشاند نور و سگ عوعو کند هر کسی بر طینت خود می‌تند	ز حرف حق لب از آن بسته‌ام که چون منصور حدیث راست مرا دار می‌شود چه کنم	خفتگان را خبر از محنت بیداران نیست تا غمت پیش نیاید غم مردم نخوری	منصور سر گذاشت در این راه و برنگشت زاهد در این غم است که دستار می‌رود	4
بگفتا جان‌فروشی در ادب نیست بگفت از عشق‌بازان این عجب نیست	جان فدای شکر شیرین شورانگیز او کز فراقش دست بر سر چون مگس باشد مرا	در سر من نیست الا وصل آن دلبر هوس تا سرم بر جای باشد این هوس باشد مرا	خواهم افکندن ز دست دل سر اندر پای دوست گر ز من بپذیردش این فخر بس باشد مرا	بر وصالش یک نفس گر دسترس باشد مرا حاصل عمر عزیز آن یک نفس باشد مرا	3

Fig. 1. Screenshot of MCQs in Excel format.

It must be noted that Doc2Vec by itself is not designed for the task of MCQ answering. However, as our current task deals with questions that require a test-taker to discern the thematic similarity between verses, sentence embeddings generated by Doc2Vec are believed to acceptably represent the meaning of documents, and documents dealing with the same topic frequently have similar embeddings. It is for this reason that we believe sentence embeddings generated by feeding each verse into a trained Doc2Vec model can be utilized to answer these specific questions.

In order obtain accurate vector representations, we trained a Doc2Vec model on a corpus of Persian poems collected from the Ganjoor website [12], which contains poems from dozens of prominent poets. This corpus is accessible on the Github website and contains a total of 8 million words of Persian poems [13]. Admittedly, the size of the corpus limits the model's capability to produce representative sentence embeddings, and expanding the size of the corpus could yield better results.

The model was trained using the library provided by Genism [14]. In order to train the model, the vector size was set to 50, and the number of epochs was set to 40. Subsequently, the trained model was used to obtain vector representations from the verses of the stem and the options. These vectors were then compared based on cosine similarity, and the option verse with the highest similarity to the stem verse was selected as the correct answer by the model. Finally, answers were compared to the answer key to determine the accuracy of the model. A flowchart describing the steps taken in this experiment is presented in Fig. 2.

V. Results and Discussion

The model managed to attain an accuracy of %38, which is a %6 improvement over the previous benchmark (Table I). By comparing the results obtained from this experiment with that of [3], it can be observed that out of all the questions which the two models answered correctly, only 12 questions were answered correctly by both models. This difference can be noteworthy, as the ultimate task is to find a model whose performance best correlates with that of students, not necessarily obtain better results.

TABLE I
Performance of the Models

Model	Accuracy
Doc2Vec	%38
BERT	%32
Random Guess	%25

By comparing the questions which both BERT and Doc2Vec answered incorrectly, we can see that out of 43 such questions, only in 14 instances (%32.5) have the models selected the same answer. Considering that there are three incorrect options for each question, no meaningful behavior while answering questions incorrectly can be discerned for either of the models.

One of the important advantages of the Doc2Vec model is its performance on questions which required interpretation and could not be answered based on semantic similarity. While the pre-trained multilingual BERT model had answered these questions randomly (%25), the Doc2Vec model has managed to correctly answer %46 of such questions.

Despite this improvement, the model's performance when facing questions that involved semantic similarity is slightly poorer. These questions usually contained incorrect options which were semantically similar to the stem and were intended to distract test-takers. The Doc2Vec model's performance on these questions was %27.27, which was slightly lower than the %30.3 performance of BERT on such questions (Fig. 3).

The fact that Doc2Vec was trained on a corpus of Persian poems may have made it more vulnerable to distractors, but it has probably also made its incorrect answers more representative of the answers of an actual student. This hypothesis will, however, need to be validated in future studies.

VI. Conclusion

This experiment was conducted to test the ability of a Doc2Vec model to answer thematic similarity MCQs when

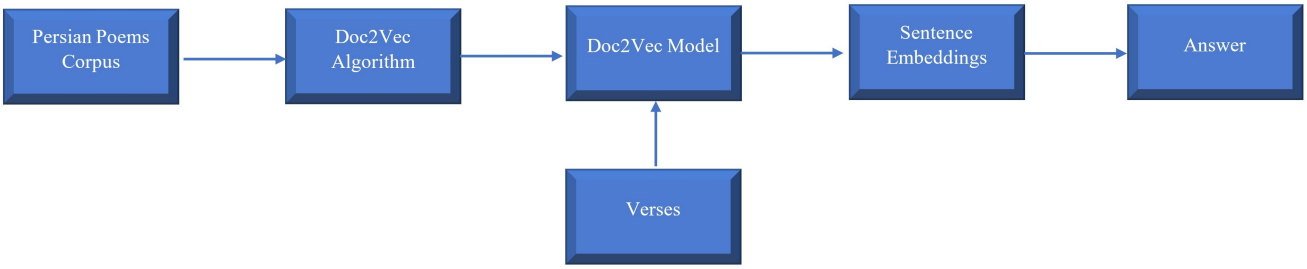

Fig. 2. Experiment flowchart.

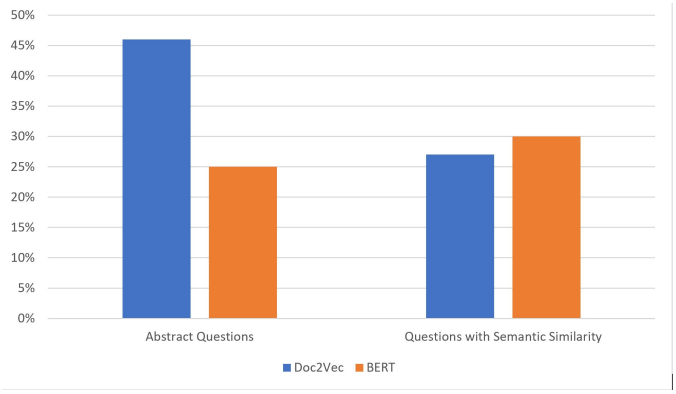

Fig. 3. Performance of Doc2Vec and BERT on Two Types of Questions.

trained on a corpus of Persian poems. The model managed to attain a %6 improvement over the previous benchmark, which had used the pre-trained multilingual BERT model to answer these questions.

As the ultimate goal of this task is to develop a model whose performance would correlate with that of actual students, the fact that Doc2Vec and BERT had little in common in terms of how they had answered the questions allows us to select the model that best resembles the performance of a student in our future works.

One way that Doc2Vec differed with BERT was its ability to answer more abstract questions, which require an interpretation of the verses. Furthermore, the fact that Doc2Vec was more prone to incorrectly answer questions that contained semantic distractors could potentially make its behavior similar to actual students.

In the next steps, we plan to fine-tune the multilingual BERT model using the same corpus of Persian poems used in this experiment, which we hope would further improve the system's accuracy. Moreover, we plan to determine which model's performance correlates better with the performance of human test-takers by obtaining a dataset of thematic similarity MCQs which contains human test-takers' performance for each individual question. Ultimately, the final model's performance will be used to estimate the difficulty of these questions.

REFERENCES

[1] L. A. Ha, V. Yaneva, P. Baldwin, and J. Mee, "Predicting the difficulty of multiple choice questions in a high-stakes medical exam," In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2019.
[2] A. Shojaei and M. Nazari, "Complete topical literature [Adabiat-e mozueeye kamel]," Tehran: Gaj International Publications, 2019.
[3] S. Akef and M. Bokaei, "Answering poetic verses' thematic similarity multiple-choice questions with BERT," In: 28th Iranian Conference on Electrical Engineering, 2020.
[4] C. Schoenick, P. Clark, O. Tafjord, P. Turney, and O. Etzioni, "Moving beyond the Turing Test with the Allen AI Science Challenge," 2016.
[5] P. Rajpurkar, J. Zhang, K Lopyrev, and P. Liang "SQuAD: 100,000+ questions for machine comprehension of text," In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016.
[6] Z. Zhang, J. Yang, and H. Zhao, "Retrospective reader for machine reading comprehension," 2020.
[7] L. A. Ha, V. Yaneva, "Automatic question answering for medical MCQs Can it go further than information retrieval," 2019.
[8] E. Asgari and J. Chappelier, "Linguistic resources & topic models for the analysis of Persian poems," 2nd Work. Comput. Linguist. Lit. (CLfL 2013), no. Ic, pp. 23–31, 2013.
[9] A. Rahgozar and D. Inkpen, "Bilingual Chronological Classification of Hafez's Poems," pp. 54–62, 2016.
[10] Q. Le and T. Mikolov, "Distributed representation of sentences and documents," 2014.
[11] A. Rahgozar and D. Inkpen, "Semantics and Homothetic Clustering of Hafez Poetry," pp. 82–90, 2019.
[12] 'Ganjoor'. [Online]. Available: http://ganjoor.ir/. [Accessed: 15-November- 2020].
[13] A. Ghaderi, 'Persian poems corpus', 2019. [Online]. Available: https://github.com/amnghd/Persian_poems_corpus/. [accessed 15-November- 2020].
[14] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010.