IJICTR
International Journal of Information & Communication Technology Research

Volume 12- Number 2 – 2020 (46 -53)

Thematic Similarity Multiple-Choice Question Answering with Doc2Vec: A Step Toward Metaphorical Language Processing

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

Received: 7 October 2019 - Accepted: 26 January 2020

Abstract—**This paper reports our improvement over the previous benchmark of the task of answering poetic verses' thematic similarity multiple-choice questions (MCQs). In this experiment, we have 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. Provided that a large-scale thematic similarity MCQ dataset is developed, the performance of a language representation model on this task could be considered as a novel benchmark to measure the capacity of a model to understand metaphorical language.**

Keywords—Doc2Vec; MCQ answering; computational linguistics; poetry; figurative speech; digital humanities.

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 prohibitively laborsome and time-consuming task, which is traditionally achieved by piloting the test item on a small sample of students.

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

* Corresponding Author

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.

The significance of these questions, however, goes beyond their prevalence in the Iranian university entrance exam. The benchmark used to evaluate language representation models is usually those models' performance on downstream tasks, such as the question answering task introduced by [2]. However, this question answering task is merely a reading comprehension task, which challenges a model to find the answer to a given question in a given text or refrain from answering if the answer cannot be found within that text. As the most recent language representation models have achieved above-human and near-perfect performance results on this task, a trend has begun to emerge to create more challenging question answering datasets, such as [3], which contains questions that require an understanding of social relations. As current language representation models lack the ability to understand metaphorical language, the performance of a language representation model on the task of poetic verses MCQ answering could be considered as a benchmark for that model's ability to understand metaphorical language. Despite the current models' inability to interpret metaphorical language, it has been shown that with the help of sentence embeddings generated using the pre-trained multilingual BERT model, an intelligent system would be able to attain an accuracy of 32%, which was a 7% improvement over the random guess baseline [4]. In the present work, we have attempted to improve on that performance by training a Doc2Vec model on a corpus of Persian poems.

The challenges of the current task are twofold: answering MCQs and processing Persian poems. While the task of question answering by itself poses certain challenges, MCQ answering could be considered an even more challenging task, as the correct answer must be selected relative to the other options. Provided that the options of a particular question are similar, the ability to distinguish the preferable answer could be a nuanced task for an intelligent system whereas a range of answers could be deemed acceptable for the task of question answering.

To add further complications to the natural language processing (NLP) aspect of this experiment, the current task does not deal with questions containing everyday language but rather questions containing poetic verses. Due to the lax rules of syntax in poetry as opposed to prose, the use of infrequent words or different connotations of words, and lack of open-source preprocessing tools for Persian poems; processing Persian poems is a much more challenging task than processing Persian prose.

In the subsequent sections, previous efforts concerning MCQ answering and Persian poem 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.

II. Related Work

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 [5].

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 Stanford 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 [2], while the best current model has achieved an F1 score of over 0.93.

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) [5].

As a result of the increased interest in question answering tasks requiring commonsense knowledge, a dataset containing more than 12000 questions which required commonsense knowledge was created by [6] through crowdsourcing. Fine-tuning the state-of-the-art language representation BERT-Large on this dataset yielded an accuracy of 56%, which is considerably lower than the human performance of 89%.

Moreover, attempts have been made to answer medical exam MCQs without the aid of any MCQ training data. In order to tackle this challenge, an artificial neural network was trained on a dataset of medical papers, with the abstract as the input and the title of the paper as the target value. Subsequently, in order to answer the questions, the question stem was fed to the model as input in place of the abstract, and the title received as the output was regarded as the correct answer. An accuracy of 39.6% was obtained on a dataset of six-option MCQs. by combining the neural network approach with information retrieval [1].

In order to develop a language representation model capable of understanding commonsense knowledge, [7] has taken advantage of knowledge bases to fine-tune BERT. They proceeded to evaluate their model on four

question answering datasets containing questions which required commonsense knowledge and obtained the best results on three of these four datasets.

Nonetheless, a model that can answer some commonsense questions does not necessarily perform well on all kinds of commonsense questions. For instance, [3] created a dataset containing 38000 three-option MCQs which required an understanding of social relations among humans in order to answer. These questions described a social situation and prompted the test-taker to answer questions with regard to the motivation behind actions, possible future events, and emotional reactions by people. Initial results demonstrated a gap of 20% between human performance and machine performance on these questions.

A need for novel and more challenging datasets is also felt in areas other than question answering. For instance, [8] argues that current neural network models have attained an accuracy of 90% on a pronoun resolution dataset, which was previously considered impossible to answer for statistical models and has therefore created a pronoun resolution dataset which would require commonsense knowledge in order to answer.

Extending the results of [9], the current paper is an improvement over [4], 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.

Persian Poem Processing

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

The first major attempt to take advantage of NLP techniques to analyze Persian poems was [10]. This work attempted to cluster approximately 18,000 Persian ghazals by 30 different poets using probabilistic topic modeling and concluded that a latent Dirichlet allocation (LDA) model achieved the best result with approximately 500 topics.

Furthermore, [11] has used the results of [10] to make conclusions about the interpretive unity of ghazal poetry, which has been the topic of some debate among literary scholars.

In another work, which also utilized probabilistic topic modeling, Hafez's ghazals were classified chronologically using a support vector machine (SVM) classifier [12]. Subsequently, the features used in this work were expanded by introducing word embeddings and other innovative features in order to cluster Hafez's ghazals [13].

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. In this experiment, we have used two distinct thematic similarity MCQ datasets. The first dataset is referred to as the Gaj dataset, as it was manually extracted from one of Gaj Publication's supplementary books. This is the same dataset used in [4], and using this dataset makes a comparison between the results possible. The second dataset used in this experiment is referred to as the Ghalamchi dataset, as it was gathered from Ghalamchi Organization's mock tests. Each of these datasets, as well as the Persian poems corpus used to train the Doc2Vec model, is described in detail in the following sections.

Gaj Dataset

The Gaj dataset contains 100 thematic similarity MCQs stored in a tabular format with the first column containing the stem verse and the second to fifth columns containing option verses. The correct answer to each question is stored in a separate column.

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, only the so-called type-one questions, which require a test-taker to select the option most similar to the stem, were included in this dataset. A sample of the dataset is presented in Fig. 1.

By analyzing the questions, it was observed that 32% of the questions in this dataset lacked semantic

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 براتی که نوشتند تغییر نیاید که ز دیوان الست است	بیشی 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 گیمیش به دست است گر زین دو فزون می‌طلبد آزیرست 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 کلهی 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 است	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

Figure 1. Sample of Gaj dataset in a tabular format.

hints and constituted abstract verses. These questions contained highly metaphorical language, which current language representation models are not expected to understand.

In the other 68% of the questions, semantically similar words could be observed between the stem verse and the option verses. However, it is worth noting that semantic similarity by itself is not enough to answer these questions, as many incorrect options contain semantically similar words to the stem in order to distract the test-taker and make the question more challenging.

The word cloud of the stem and option verses in this dataset is presented in Fig. 2.

Ghalamchi Dataset

The Ghalamchi dataset is different from the Gaj dataset, inasmuch as it contains not only type-one questions, in which a student must select the option containing the verse most similar to the stem but also type-two thematic similarity MCQs, which prompt the test-taker to select the option which is thematically most different from the stem verse.

Moreover, as these questions were designed by different exam creators than the questions in the Gaj dataset, some variation between the performance of the model on the two datasets is expected.

This dataset contains a total of 79 questions, including 42 type-one and 37 type-two questions. Furthermore, in 45 questions (i.e., in approximately 57% of the questions), semantic similarity was observed between the verses while 34 questions lacked semantic similarity and contained more abstract poetic verses. The higher percentage of questions containing abstract verses could potentially make the Ghalamchi dataset more challenging for a language representation model than the Gaj dataset.

Out of the 42 type-one questions, 28 questions (i.e., approximately 67% of type-one questions) contain semantic similarity while only 14 questions are more abstract. Nonetheless, out of the 37 type-two questions, only 17 questions (i.e., approximately 46% of type-two questions) contain semantic similarity while 20 questions are more abstract. The distribution of the questions in the Ghalamchi dataset with regard to type and semantic similarity is presented in Table I.

The word cloud of the verses in this dataset is also presented in Fig. 3.

In both datasets, 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.

TABLE I. GHALAMCHI DATASET QUESTION DISTRIBUTION

	Type 1	Type 2	Total
Abstract	14	20	34
Semantically Similar	28	17	45
Total	42	37	79

Persian Poems Corpus

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 poems would, in theory, yield better results.

The Persian poems corpus used in this experiment is a Persian poems corpus crawled from the Ganjoor website and published on Github. This corpus contains poems by 48 stylistically diverse Persian poets who lived in different eras. Three versions of this corpus have been published: the original version, the normalized version, and the version without stop words.

This corpus contains a total of 1211277 hemistichs and a total of 7888045 tokens, 14996 of which are unique words. The word cloud of this corpus is presented in Fig. 4.

IV. METHOD

In order to answer thematic similarity MCQs, we first trained a Doc2Vec model on a corpus of Persian poems. The Doc2Vec language representation model

Figure 2. Word cloud of poems in Gaj dataset.

Figure 3. Word cloud of poems in Ghalamchi dataset.

Figure 4. Word cloud of Persian poems corpus

was introduced by [14] and was based on the previous Word2Vec model [15]. The objective of the Word2Vec model was to represent words in a vector space in which semantically similar words are close to each other. Moreover, relationships between words are also captured in this representation model, with for instance Iran and Tehran having the same relationship as France and Paris. Doc2Vec is an improvement over Word2Vec, inasmuch as it allows documents to be represented as vectors.

There are two algorithms used in the training process of the Word2Vec representation: the continuous bag-of-words model (CBOW) and the skip-gram model. In the CBOW model, using the surrounding words of a masked word, the model attempts to predict the masked word. The skip-gram model works in reverse, as it attempts to predict the surrounding words using only one word.

The innovation of Doc2Vec was adding another vector representing a document to the word vectors of that document. This vector would represent that document after training and would have all the qualities of word embeddings generated using Word2Vec, i.e., the vectors of similar documents would be closer to each other according to distance metrics as well.

In this experiment, each hemistich in the Persian poems corpus is considered a document, and therefore by training the Doc2Vec model on this corpus, we attempted to create a representation model in which thematically similar verses might also have similar sentence embeddings. Subsequently, the trained model was used to obtain vector representations, (commonly referred to as embeddings) for the verses of the stems and the options respectively. In the final stage, these embeddings were compared using cosine similarity in order to determine the correct answer.

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 in order to answer these specific questions.

In order to obtain accurate vector representations, we trained a Doc2Vec model on the normalized version of the Persian poems corpus. 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 [16]. The model was trained in 40 epochs with the vector size of the embeddings set to 50. 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 for type-one questions, and the option verse with the lowest

similarity to the stem verse for type-two questions, 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. 5.

V. Results and Discussion

The model managed to attain an accuracy of 38% on Gaj questions, which is a 6% improvement over the previous benchmark, but an accuracy of 25% on Ghalamchi questions, which despite being a poor result, performed better than BERT, which had an accuracy of %22 on the same dataset. By comparing the results obtained from this experiment with that of [4], it can be observed that out of all the Gaj 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.

By comparing the Gaj questions which both BERT and Doc2Vec answered incorrectly, we can observe 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 Gaj questions which required an interpretation of the verses 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.

When answering Ghalamchi questions, Doc2Vec has answered 12 out of 45 questions with semantic similarity, while BERT only answered 7 of such questions correctly. However, Doc2Vec performed more poorly on abstract questions of this dataset by answering only 7 out of 34 of such questions. BERT managed to answer 10 of such questions correctly. Moreover, Doc2Vec had a much better performance on type-two questions than BERT. By answering 11 out of the 37 type-two questions and only 8 out of the 42 type-one questions, Doc2Vec demonstrated a totally different approach, as BERT had answered only 5 out of the 37 type-two questions and 12 out of the 42 type-one questions. An overview of the performance of the two models on different questions of the two datasets is provided in Table II.

While BERT is generally considered a more sophisticated language representation model than Doc2Vec and has yielded better results on other downstream tasks, a number of factors may have played a role in Doc2Vec's better performance on this task. Considering the fact that the multilingual BERT model used in [4] was trained a corpus of Wikipedia articles, the sentence embeddings generated using that model could not accurately represent poetic verses, as the syntactic structure of Persian poems and also the meaning and connotation of some words in poems are very different than those of everyday Persian. As a result, embeddings generated using a Doc2Vec model trained on an even small corpus of domain-specific text have yielded better results than embeddings generated using a more sophisticated model such as BERT trained on an unrepresentative corpus.

Another factor which could have played a role in Doc2Vec's better performance is the fact that generating sentence embeddings using BERT without fine-tuning is not a standard procedure, unlike using models such as Doc2Vec or fastText. Considered as a landmark in NLP, BERT's innovation was that it could achieve a better understanding of context through its unique training, which encoded sentences in a bidirectional manner [17]. While BERT and its variations have achieved state-of-the-art results on a number of tasks, the application of BERT is usually done through fine-tuning it on a labeled dataset. Generating sentence embeddings using BERT is not as straightforward as it is using the aforementioned language representation models, and even then, there is some debate about their effectiveness. In order to fully unleash BERT's potential for this task, a large-scale dataset of thematic similarity MCQs needs to be developed, and by fine-tuning BERT on that MCQ dataset, one can expect better results.

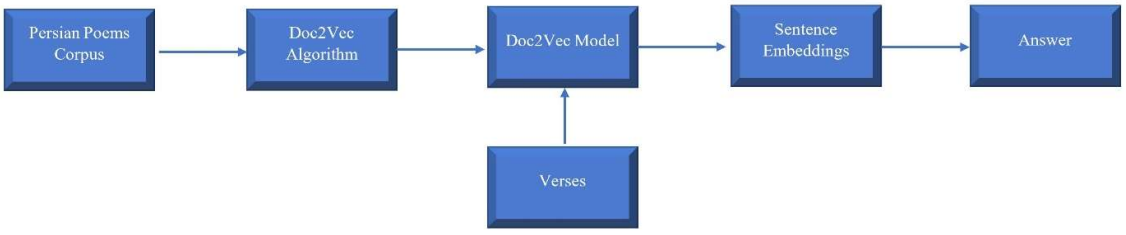

Figure 5. Experiment flowchart

TABLE II. PERFORMANCE OF THE MODELS ON DIFFERENT QUESTIONS OF GAJ AND GHALAMCHI DATASETS

	Abstract Ghalamchi	Semantic Similarity Ghalamchi	Ghalamchi Type 1	Ghalamchi Type 2	Abstract Gaj	Semantic Similarity Gaj	Ghalamchi Total	Gaj Total
Doc2Vec	21%	27%	19%	30%	46%	27%	25%	38%
BERT	26%	16%	29%	14%	25%	34%	22%	32%

VI. CONCLUSION

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

This paper has, for the first time, introduced a task which could be used to evaluate a language representation model's ability to understand and interpret figurative or metaphorical language (i.e., language that intends to convey a different meaning than the one denoted by the literal meaning of the words). Ambitious as this feat may seem, the development of a means to measure progress is often the first step toward progress. Furthermore, considering the fact that figurative language exists in virtually all languages, the development of such a dataset is possible for languages other than Persian. However, as thematic similarity MCQs may not be as prevalent in other educational systems as they are in Iranian education, such datasets could be developed through crowdsourcing.

The current paper has also introduced a novel way of answering MCQs, which is utilizing the cosine similarity between the embeddings of the options of an MCQ.

Absent a large MCQ training dataset, the cosine distance between the sentence embeddings generated from the text of the question stem and the question options may be used in order to answer MCQs. This strategy, however, only works for questions where the similarity between the options could be exploited in order to answer the questions, and other strategies need to be developed in order to answer other kinds of MCQs where similarity is not a deciding factor.

This experiment has paved the way for the development of a large-scale thematic similarity MCQ dataset, which would serve as a benchmark for evaluating the ability of a language representation model to understand metaphorical speech and writing, which is an extremely challenging task. Moreover, determining which model's performance correlates better with the performance of human test-takers could be a worthwhile task.

Furthermore, we have demonstrated that despite the relatively small size of the Persian poems corpus with less than 8 million words, sentence embeddings generated using this corpus are meaningful and can therefore be used in future research involving Persian poems.

REFERENCES

[1] L. A. Ha and V. Yaneva, "Automatic question answering for medical MCQs: Can it go further than information retrieval?" in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria: INCOMA Ltd., Sep. 2019, pp. 418–422. [Online]. Available: https://www.aclweb.org/anthology/R19-1049

[2] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuad: 100,000+ questions for machine comprehension of text," *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, no. ii, pp. 2383–2392, 2016.

[3] M. Sap, H. Rashkin, D. Chen, R. Le Bras, and Y. Choi, "Social IQA: Commonsense reasoning about social interactions," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 4463–4473, 2020.

[4] S. Akef and M. H. Bokaei, "Answering poetic verses' thematic similarity multiple-choice questions with bert," in *28th Iranian Conference on Electrical Engineering*, 2020.

[5] C. Schoenick, P. Clark, O. Tafjord, P. Turney, and O. Etzioni, "Moving beyond the turing test with the allen AI science challenge," *Communications of the ACM*, vol. 60, no. 9, pp. 60–64, 2017

[6] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "CommonsenseQA: A question answering challenge targeting commonsense knowledge," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4149–4158. [Online]. Available: https://www.aclweb.org/anthology/N19- 1421

[7] A. Mitra, P. Banerjee, K. K. Pal, S. Mishra, and C. Baral, "How Additional Knowledge can Improve Natural Language Commonsense Question Answering?" arXiv e-prints, p. *arXiv*:1909.08855, Sep. 2019.

[8] K. Sakaguchi, R. Le Bras, C. Bhagavatula, and Y. Choi, "Winogrande: An adversarial winograd schema challenge at scale," *arXiv*, 2019.

[9] S. Akef, M. H. Bokaei and H. Sameti, "Training Doc2Vec on a Corpus of Persian Poems to Answer Thematic Similarity Multiple-Choice Questions," *2020 10th International Symposium on Telecommunications (IST)*, 2020, pp. 146-149, doi: 10.1109/IST50524.2020.9345918.

[10] E. Asgari and J. Chappelier, "Linguistic resources & topic models for the analysis of Persian poems," *2nd Workshop on Computational Linguistics for Literature* (CLfL 2013), no. Ic, pp. 23–31, 2013. [Online]. Available: http://www.aclweb.org/anthology/W13-1404

[11] E. Asgari, M. Ghassemi, and M. A. Finlayson, "Confirming the themes and interpretive unity of Ghazal poetry using topic models," *Proceedings of the NIPS Workshop on Topic Models: Computation, Application, and Evaluation*, p. Submission 18, 2013. [Online]. Available: http://mimno.infosci.cornell.edu/nips2013ws/nips2013tm submission 18.pdf

[12] A. Rahgozar and D. Inkpen, "Bilingual chronological classification of hafez's poems," in *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*. San Diego, California, USA: Association for Computational Linguistics, Jun. 2016, pp. 54–62. [Online]. Available: https://www.aclweb.org/anthology/W16-0207

[13] ——, "Semantics and homothetic clustering of hafez poetry," in *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Minneapolis, USA: Association for Computational Linguistics, Jun. 2019, pp. 82–90. [Online]. Available: https://www.aclweb.org/anthology/W19-2511

[14] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, p. II–1188–II–1196.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[16] 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*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACLHLT*, 2019

Soroosh Akef received his B.A. in English language and literature from Ferdowsi University of Mashhad in 2018 and his M.Sc. in computational linguistics from Sharif University of Technology in 2021. His current research interests include the applications of natural language processing in literature and education.

Mohammad Hadi Bokaei received the B.Sc. degree from the Department of Computer Science of Iran University of Science and Technology, Tehran, Iran, in 2008. He received the M.Sc. and Ph.D. degrees in artificial intelligence from Sharif University of Technology, Tehran, Iran, in 2010 and 2015, respectively. He collaborated with Dr. Yang Lius at his Speech and Language Processing Lab as a visiting student in 2014. He is currently an assistant professor at Iran Telecommunication Research Center, and his current research interests include natural/spoken language processing, speech summarization, spoken dialogue systems, and machine learning.

Hossein Sameti received his Ph.D. in Electrical Engineering from University of Waterloo, Canada, in 1995. He received his M.Sc. and B.Sc. in Electrical Engineering from Sharif University of Technology, Tehran, Iran in 1989 and 1986 respectively. He is now an associate professor at the Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, where he has served as the head of Artificial Intelligence group for 6 years and as the department chair for 4 years. He is the founder and supervisor of Speech Processing Lab (SPL) in that department, too. His research areas are speech recognition and understanding, spoken dialogue systems, speaker identification and verification, speech enhancement, and natural language processing.