A Study of Word Embeddings for Biomedical Question Answering

Sanjay Kamath, Brigitte Grau, Yue Ma

A Study of Word Embeddings for Biomedical Question Answering

SANJAY KAMATH*,**, BRIGITTE GRAU*,***, YUE MA**

*LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
**LRI, Univ. Paris-Sud, CNRS, Université Paris-Saclay, France
sanjay@lri.fr, brigitte.grau@limsi.fr, yue.ma@lri.fr
***ENSIIE

Abstract. Different word embeddings for question answering task can give different impacts on the performance. This article discusses the issue, compares different word embeddings on a question answering task for biomedical domain and reports better performing word embeddings.

1 Introduction

Question Answering (QA) focuses on giving a user precise answers to the questions posed in natural language. The QA task which extracts answers from given paragraphs is known as Extractive QA. This task is sometimes referred to as Machine Reading task. Since deep learning models are used widely to address these problems, Rajpurkar et al. (2016) released a dataset, SQUAD, consisting of 100,000+ questions posed by crowdworkers on a set of Wikipedia passages. Results of several Deep Neural Network (DNN) models using the SQUAD Dataset can be found on the leaderboard [1] which show good performance.

Domain-specific extractive QA focuses on question-answer pairs restricted to a domain and is evaluated in the BIOASQ challenge Task B for the biomedical domain - Tsatsaronis et al. (2015). Biomedical questions with their exact answers, relevant text snippets, concepts, articles, summaries were constructed by biomedical experts from around Europe -Nentidis et al. (2017). An example question with one snippet from BIOASQ data is shown below.

Question:	What is the mode of inheritance of Wilson's disease?
Answer:	autosomal recessive
Snippets:	The overall sex ratio of patients was nearly 1:1, and genetic analysis of 20 families confirmed an autosomal recessive mode of inheritance.

One of the biggest challenges for domain-specific QA, such as BIOASQ, is the scarcity in the amount of data available. Deep learning based models often use large scale datasets for supervised learning built on open domain data such as wikipedia dumps, to improve the accuracy of QA systems. On the other end, biomedical domain despite of having great amount of resources such as UMLS thesaurus, ontologies such as SNOMED CT, and tools such as Metamap, lacks large scale datasets with QA examples for the question answering task. As far

1. https://rajpurkar.github.io/SQuAD-explorer/

as we know, BIOASQ dataset is the only dataset for biomedical QA, whose size is certainly small to train deep learning based models without overfitting. Hence building a deep neural network model trained only on BIOASQ data for factoid questions (approx. 400 questions) would not be suitable.

As it is impossible to create a large dataset for biomedical QA without extensive efforts of domain experts, transfer learning can be an alternative approach as used by Wiese et al. (2017). The authors train a neural network model based on FASTQA - Weissenborn et al. (2017) on open domain data using SQUAD dataset, and then use it to retrain the model on the BIOASQ dataset.

In this paper, we use a similar approach for transfer learning with the DRQA model by Chen et al. (2017) as it obtains comparable results on open domain QA and its implementation is available [2]. We perform the experiments using our annotated dataset (section 3) for BIOASQ.

It has been previously shown in several works that the usage of different word embeddings in the Input layer has different impacts on the overall performance of a DNN model. This paper presents the experiments we made by using different available pretrained embedding models along with the ones we trained using different hyperparameters for Word2Vec models, on the task of biomedical QA. We found that the Global Vectors (Glove) by Pennington et al. (2014) which is trained on open domain data (Wikipedia) performs with the best scores and also the Skipgram model with 300 dimensions on Wikipedia plus Biomedical data performed better for some test sets.

2 The model

We present here the adaptation of an existing model named DRQA reader by Chen et al. (2017) to the biomedical domain. DRQA reader has three components: 1) Input layer: where the input question words and input passage words are encoded using a pretrained word embedding space; 2) Neural layer: RNN or LSTM; 3) Output layer: or decoding layer, where the outputs are start and end tokens representing a span of an extracted answer.

In the input layer, word embeddings are used to encode the words of paragraphs and question into vectors, along with textual features such as Part of Speech tags, Named-Entity tokens, Term frequencies of the words in the paragraph. Authors use *Aligned question embeddings* where an attention score captures the similarity between paragraph words and questions words. Neural layer, where the core DNN model is defined uses different NN architectures to capture semantic similarities between the QA pairs. In the output layer, two independent classifiers use a bilinear term to capture the similarity between paragraph words and question words and compute the probabilities of each token being start and end of the answer span. An argmax value over the unnormalized exponential is calculated on the spans, to get a final prediction.

Following this model, we first train the DRQA model on SQUAD dataset with its default hyperparameters, and then retrain the model on BIOASQ questions which are relatively small in quantity. Doing so, the model would learn how to extract answer spans from open domain questions, which is later fine tuned for domain specific data. We perform the above mentioned experiment with different embedding spaces which are detailed in the Section 3 to see how well the QA system performs on different input encodings.

2. https://github.com/facebookresearch/DrQA

3 Experiments and Results

The dataset of BIOASQ for task B consists of questions, relevant snippets, and answers provided by biomedical experts. However, the answers are written manually or chosen from concepts annotation tool, thus may be not exactly a snippet segment. This makes an automatic projection of the answers in the snippets difficult. Whereas, Extractive QA models use answers extracted from the paragraph and their offsets as input to the model during training. Hence, using BIOASQ data directly for such models would result in false negatives because of the absence of answer segments in the snippets which results in absence of answer offsets for training.

On another hand, while evaluating a system for Accuracy and MRR, BIOASQ Task B considers exact string matching measure. If an answer in the gold standard is *"Transcription factor EB (TFEB)"* and the predicted answer is *"Transcription factor EB"*, it is marked as a wrong answer because of lack of the term *"(TFEB)"*, which poses a serious concern during evaluation. To overcome the issues discussed above, we manually annotated BIOASQ challenge 5B dataset for factoid questions. The annotations contain offsets and the corresponding strings which we believe are semantically correct as answers, though they may be not identical to the exact answers given by the experts.

BIOASQ Task B contains five different test batches with distinct question sets. We retrained the DRQA model on BIOASQ 2017 5B training data (with our annotations) after removing each test set. For the embeddings part, we train different word embeddings spaces with CBOW and Skipgram models with different hyperparameters and different data. We chose the BIOASQ 5A task data which consist of 12.8 Million PUBMED articles as an input corpus (referred as *BIOASQ17*). We preprocessed this dataset to remove special characters and use the Gensim tool to train word embeddings with 100, 200, 300, 400 dimensions(D). For experiments with existing pretrained embeddings, we use word vectors provided by BIOASQ task (referred as *BIOASQ16*), and Global Vectors (*Glove*).

Table 1 presents the comparison of four word embedding spaces tested on the five test sets (Test-1 to Test-5) and a set with all test sets combined (All). *BIOASQ16* with 200D performed worse on our experiments. *BIOASQ17* were trained by us with Skipgram model and 300D. We can see that although *Glove* are trained on Web Crawl data and not specifically biomedical data, it performs better than the rest trained on biomedical data because of the large training data of Glove. *Wiki+BIOASQ17* trained with Skipgram and 300D on data of *BIOASQ17* and Wikipedia articles, has second best accuracy after *Glove* because of the domain specific training data even though it is smaller compared to Glove's training data.

Table 2 presents a comparison of different embedding spaces trained on different dimensions (namely 100, 200, 300, 400) with CBOW and Skipgram models as described in Mikolov et al., where the performance is calculated based on Strict Accuracy measure of BIOASQ 4B test sets. It is evident from the table that Skipgram performs better than CBOW when the dimensions are higher. But 300 dimensions is found to be optimal in terms of both strict and lenient accuracy. Increasing it to 400 dimensions did not fetch better results.

These experiments highlight the importance of choosing right word embeddings for biomedical domain QA system. Glove performs better on average because of large amount of data it is trained on, and pretraining on SQUAD which has a large set of open domain questions makes the pretrained QA model learn better representations. Whereas the biomedical embeddings are trained on lesser data and domain specific vocabulary which has impact over the pretraining of

	Nb. of Ques.	BIOASQ16 |V| = 1.7M |T| = 2.2B	BIOASQ17 |V| = 2.1M |T| = 2.3B	Wikipedia |V| = 2.13M |T| = 2.19B	Wiki+BIOASQ17 |V| = 4M |T| = 4.49B	Glove |V| = 2.2M |T| = 840B
Test-1	39	0.3333	0.4615	0.4615	0.5128	**0.5897**
Test-2	31	0.3548	0.4839	0.4516	0.4838	**0.5161**
Test-3	26	0.3846	**0.6538**	**0.6538**	**0.6538**	0.6153
Test-4	31	0.3548	0.4516	**0.5161**	0.4193	0.4193
Test-5	33	0.4545	0.5757	0.6061	0.6061	**0.6666**
Average	–	0.3764	0.5253	0.5378	0.5352	**0.5614**
All	160	0.3312	0.5125	0.45	0.5	**0.525**

TAB. 1 – *Accuracy (top 5) on 4B test with different Embeddings: |V|= vocab,|T|= token counts*

	CBOW		Skipgram	
Dims	Strict	Lenient	Strict	Lenient
100	0.3062	**0.4875**	0.3125	0.5
200	0.2875	0.475	**0.3375**	0.5
300	**0.3187**	**0.4875**	0.3187	**0.5125**
400	0.2875	0.4687	0.3	0.4875

TAB. 2 – *Comparision of Word2vec models on 4B Test set (Testset "All" from Table 1)*

SQUAD. Hence we plan to investigate different embeddings in different phases of training for future work and we will explore to learn embeddings with a bigger mixed corpus.

References

Chen, D., A. Fisch, J. Weston, and A. Bordes (2017). Reading wikipedia to answer open-domain questions. In *Proceedings of ACL 2017*, pp. 1870–1879.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space (2013). arxiv preprint. *arXiv preprint arXiv:1301.3781*, 1532–1543.

Nentidis, A., K. Bougiatiotis, A. Krithara, G. Paliouras, and I. Kakadiaris (2017). Results of the fifth edition of the bioasq challenge. In *BioNLP 2017*, pp. 48–57.

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pp. 1532–1543.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP 2016*, pp. 2383–2392.

Tsatsaronis, G., G. Balikas, P. Malakasiotis, et al. (2015). An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics 16*(1), 138.

Weissenborn, D., G. Wiese, and L. Seiffe (2017). Making neural qa as simple as possible but not simpler. In *Proceedings of CoNLL 2017*, pp. 271–280.

Wiese, G., D. Weissenborn, and M. Neves (2017). Neural domain adaptation for biomedical question answering. In *Proceedings of CoNLL 2017*, pp. 281–289.