JU_NITM at IJCNLP-2017 Task 5: A Classification Approach for Answer Selection in Multi-choice Question Answering System

Sandip Sarkar
Computer Science and Application
Hijli College, Kharagpur
sandipsarkar.ju@gmail.com

Dipankar Das
Computer Science and Engineering
Jadavpur University, Kolkata
dipankar.dipnil2005@gmail.com

Partha Pakray
Computer Science and Engineering
NIT Mizoram, Mizoram
parthapakray@gmail.com

Abstract

The present task describes the participation of the JU_NITM team in IJCNLP-2017 Shared Task 5: ”Multi-choice Question Answering in Examinations”. One of the main aims of this shared task is to choose the correct option for each of the multi-choice questions. We represent each of the questions and its corresponding answer in vector space and find the cosine similarity between two vectors. Our proposed model also includes supervised classification techniques to find the correct answer. Our system was only developed for the English language, and it also obtains an accuracy of 40.07% for the test dataset and 40.06% for validation dataset, respectively.

1 Introduction

In the era of computer and internet, we are getting any information at our fingertips. Besides, Natural Language Processing (NLP) is used to improve computer intelligence and to understand the natural languages given by us. If we consider information retrieval, semantic level of matching also holds a major role to retrieve similar documents (Onal et al., 2016). Question Answering (QA) as a sub field of Natural Language Processing (NLP) and Information Retrieval (IR) aims to give answer of a given question in natural language. In the present task, we proposed a method to choose correct answer in multi-choice question answering domain. First, we find the relation between the question and answers in the same vector space using cosine similarity. After that, we used a supervised classification approach to predict the correct answer from multiple options that are considered as our classes.

The rest of the paper is organized as follows. Section 2 describes the detail information of this shared whereas related work is discussed in Section 3. Similarly, Section 4 describes our proposed model in detail. Dataset information of this shared task is given in Section 5. We present results of our system in Section 6. Finally, Section 7 presents the conclusions and future work.

2 Task Overview

IJCNLP-2017 Task 5: ”Multi-choice Question Answering in Examinations” challenged the participants to automatically predict the correct answer in multi choice Question Answering in exams. This shared task contains complex questions like in real exam. All questions are from the elementary and middle school level. Each question contains four possible answers. Questions are collected from different domains like biology, physics, chemistry etc. The format of the dataset is given in Table 1. In Table 1 in Correct answer column 0,1,2,3, denote answer A, answer B, answer C, answer D respectively.

3 Related Work

In recent time community question answering (cQA) plays an important role to find desired information. Many researchers proposed different type of approaches to deal with community question answering. To solve this problem most of the researcher used traditional information retrieval system.

Besides the use of neural network in information retrieval is gradually increasing because of their advantages. In the same time, distributed semantic model plays an important role to find the similarity between two sentences or documents (Sarkar et al., 2016a). Neural language model for learning distributed vector representations of

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 213–216,
Taipei, Taiwan, November 27 – December 1, 2017. ©2017 AFNLP

Quesion	Answer A	Answer B	Answer C	Answer D	Correct answer
What component of blood carries oxygen?	red blood cells	white blood cells	plasma	platelets	0
The primary component of steel is _____.	copper	iron	cobalt	nickel	3
The deepest canyon in the ocean floor.	Hellenic Trench	Philippine Trench	Marianas Trench	Japan Trench	2
Which are examples of altricial birds?	the domestic chicken	ducks	the magapodes	the Great Frigatebird	3
The nucleus contains	protons	neutrons	electrons	two of the above	3

Table 1: Multiple Choice Question With Correct Answer

words is known as word embedding. The name of those methods are the continuous bag-of-words model and the skip-gram model. These methods capture higher order relationships between words and sentences. SemEval- 2015 Task 3 on Answer Selection in cQA is similar type of shared task (Agirre et al., 2015). The main aim of SemEval-2015 Task 3 shared task was to develop a system that automatically detects the most relevant answers from the irrelevant ones.

4 System Framework

To deal with this shared task we use word embeddings to find the relation between questions and answer options. Word embeddings is well known approach for semantic textual similarity, question answering and information retrieval system(Řehůřek and Sojka, 2010; Elman, 1990). For classification, we used Matlab toolkit. [1] classification module have been used with respect to each of the corresponding runs submitted by our team to the shared task.

In this shared task, we build a complex decision tree classifier using word2vec [2] feature to predict the correct answer. Figure 1 describes our system architecture.

4.1 Distributed Semantic Similarity

Distributional semantic is very useful to capture the textual similarity between sentences. The model is mainly based on one hypothesis that the meaning of a word depends on the surrounding words. (Pennington et al., 2014) The underlying

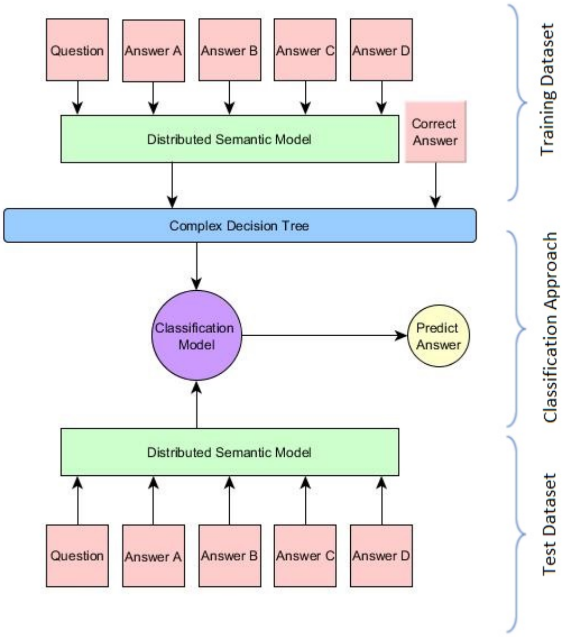

Figure 1: System Framework

[1] http://in.mathworks.com/help/stats/classification-trees-and-regression-trees.html

[2] https://radimrehurek.com/gensim/models/word2vec.html

idea of this concept is that ”a word is characterized by the company it keeps”. (Firth, 1957) Researchers are trying to improve this model that can be achieved from integrating distributional vectors semantics which is also known as word embeddings. Two such methods are the continuous bag-of-words model and the skip-gram model. These methods have been shown to produce embeddings that capture higher order relationships between words that are highly effective in natural language processing tasks involving the use of word similarity and word analogy. (Zuccon et al., 2015)

To find the relation between question and options we used the GoogleNews vectors dataset which is available on Google word2vec website. [3] The vocabulary size of this trained model is 3,000,000 word and the size of word vector is 300-dimensional. The model is trained on 100 billion words.

4.2 Classification Approach

Classification technique is used in different scientific fields to build a classification models from features of input data set. Different types of classification approach have been proposed like probabilistic neural network, rule based classifier, decision tree classifier, support vector machine to solve different types of problem (Sarkar et al., 2016b). In decision tree classification approach a series of question are asked and each time an answer is given to make a decision. Finally, we derived a conclusion of the problem. The series of question and answer are organized in the form of hierarchical structure.

For our experiment, we used complex decision tree approach to predict the correct answer. The training dataset set is used to build a classification model, which is used to predict the class labels of test dataset. The accuracy of a classification model is calculated using the count of correct and incorrect prediction by the model. .

5 Dataset

The 2017 IJCNLP Task 5 shared task collect questions from different subjects i.e. biology, chemistry, physics, earth science and life science. The variety of the questions is very challenging for the participants. The questions are given in two languages English and Chinese. We are participating only English dataset. The statistics of IJCNLP English dataset is described in the Table 2. In this shared task questions are provided in csv file where each row defines each question.

English Subset				
	Train	**Valid**	**Test**	**Total**
Biology	281	70	210	561
Chemistry	775	193	581	1549
Physics	299	74	224	597
Earth Science	830	207	622	1659
Life Science	501	125	375	1001
English Total	2686	669	2012	5367

Table 2: Statistics of IJCNLP Dataset

6 Results

In this section, we discuss about our experiment results on valid and test dataset. In the same time, we also shown the comparison between winner score and our system score on those datasets. Table 3 shows that our model gives better result with compare to winner score. Our model is not based on traditional information retrieval system. Besides our model is simple and easily implemented on different types of dataset. However, our system face problem to capture the semantic meaning of chemical equations as well as integer values.

The accuracy is calculated using the following equation

$$Accuracy = \frac{numberofcorrectquestions}{totalnumberofquestions} \quad (1)$$

Organizer implemented simple retrieval based method as a baseline, and they used Apache Lucene which is a well-known software for information retrieval. For baseline system organizer, concatenate question with option and generate the query. Next use this query to find the relevant documents. In equation 2 the query and document are denoted by q and d respectively similarity between query and document is calculated using *Sim(q,d)*. Finally, they are taking top three similarity to calculate the score using following equation

$$Score(q,a) = \frac{1}{n}\sum_{1}^{n} Sim(q,d) \quad (2)$$

7 Conclusion and Future Work

In this paper we present distributed semantic model on IJCNLP-2017 Task 5 dataset in ”Multi-choice Question Answering in Exams” shared

[3] https://code.google.com/archive/p/word2vec/

	Dataset	
	Valid	**Test**
Winner Score	0.487	0.456
JU_NITM Score	0.407	0.406

Table 3: Comparison between Winner Score and Our System Score

task. Our proposed method achieved good result compare with winner score. The advantage of distributed semantic model is that this model is simple and robust. This model not only find exact terms from the questions and answers from resources but also find semantic information from resources. Dealing with IJCNLP dataset we observe that our proposed method can be easily implemented into complex applied systems. In the same time, we face some problem to deal with chemistry dataset because our model does not represent the chemical equations in vector space.

Our future aim is to overcome from this problem. We are trying to improve our dataset from which we build our distributed semantic model so that we can represent chemical equation in vector space. In the same this we are also trying to implement doc2vec model to deal with complex system.

Acknowledgments

This work presented here is under the Research Project Grant No. YSS/2015/000988 and supported is by the Department of Science & Technology (DST) and Science and Engineering Research Board (SERB), Govt. of India. Authors are also acknowledges the Department of Computer Science & Engineering of National Institute of Technology Mizoram, India for providing infrastructural facilities and support.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science*, 14(2):179–211.

J. R. Firth. 1957. A synopsis of linguistic theory 1930-55. 1952-59:1–32.

Kezban Dilek Onal, Ismail Sengor Altingovde, Pinar Karagoz, and Maarten de Rijke. 2016. Getting Started with Neural Models for Semantic Matching in Web Search. *CoRR*, abs/1611.03305.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Sandip Sarkar, Partha Pakray, Dipankar Das, and Alexander Gelbukh. 2016a. Regression Based Approaches for Detecting and Measuring Textual Similarity. In *Mining Intelligence and Knowledge Exploration: 4th International Conference, MIKE 2016*, pages pp. 144–152, Mexico City. Springer International Publishing.

Sandip Sarkar, Saurav Saha, Jereemi Bentham, Partha Pakray, Dipankar Das, and Alexander Gelbukh. 2016b. NLP-NITMZ@DPIL-FIRE2016: Language Independent Paraphrases Detection. In *Shared task on detecting paraphrases in Indian languages (DPIL), Forum for Information Retrieval Evaluation (FIRE)*, pages pp. 256–259, Kolkata, India.

Guido Zuccon, Bevan Koopman, Peter Bruza, and Leif Azzopardi. 2015. Integrating and Evaluating Neural Word Embeddings in Information Retrieval. In *Proceedings of the 20th Australasian Document Computing Symposium*, ADCS '15, pages 12:1–12:8, New York, NY, USA. ACM.