MappSent at IJCNLP-2017 Task 5: A Textual Similarity Approach Applied to Multi-choice Question Answering in Examinations

Amir Hazem[1] **Basma El Amal Boussaha**[1] **Nicolas Hernandez**[1]
[1] LS2N - UMR CNRS 6004, Université de Nantes, France
{Amir.Hazem,Basma.Boussaha,Nicolas.Hernandez}@univ-nantes.fr

Abstract

In this paper we present MappSent, a textual similarity approach that we applied to the multi-choice question answering in exams shared task. MappSent has initially been proposed for question-to-question similarity (Hazem et al., 2017). In this work, we present the results of two adaptations of MappSent for the question answering task on the English dataset.

1 Introduction

Question-Answering is certainly one of the most challenging area of research of information retrieval (IR) and natural language processing (NLP) domains. If many investigations and countless approaches have been proposed so far, the developed systems still have difficulties to deal with text understanding. Mainly because of the complexity of the language in terms of lexical, semantic and pragmatic representations. However, with the boom of neural networks, various deep learning approaches ranging from a word level embedding representation (Bengio et al., 2003; Collobert and Weston, 2008; Mikolov et al., 2013; Pennington et al., 2014) to a longer textual level embedding representation such as phrases, sentences, paragraphs or documents (Socher et al., 2011; Mikolov et al., 2013; Le and Mikolov, 2014; Kalchbrenner et al., 2014; Kiros et al., 2015; Wieting et al., 2016; Arora et al., 2017) have been proposed and have shown promising results in many applications. Not to mention other more sophisticated approaches like recurrent neural networks (RNN) (Socher et al., 2011, 2014; Kiros et al., 2015), long short-term memory (LSTM) to capture long distance dependency (Tai et al., 2015) or convolutional neural networks (CNNs) (Kalchbrenner et al., 2014) to represent sentences.

Inspired by the new textual embedding representations (Wieting et al., 2016; Arora et al., 2017), we propose in this paper to adapt MappSent (Hazem et al., 2017) an approach that have initially been developed for question pairs similarity, to the task of multi-choice question answering in examinations. Two adaptations are proposed. The first one is a direct application of MappSent to question-answer pairs, while the second one can be seen as a pivot-based approach in which questions and answers are treated separately in two different sub-spaces. Then, the correct candidate is extracted by transitivity thanks to the similarity of test question-answering pairs regarding question-answering pairs of the training corpus. If the shared task provides datasets in two languages: Chinese and English, we only deal with English.

The remainder of this paper is organized as follows. Section 2 presents the multi-choice question answering in examination task and the provided training datasets. Section 3 describes MappSent, the textual similarity approach and its two adaptations to multi-choice question answering. Section 4 is devoted to the evaluation of the different approaches. Section 5 discusses the results and finally, Section 6 presents our conclusion.

2 Task and Resource Description

The task consists of a multi-choice question challenge. Given a question, four answers are provided and the purpose is to find the correct one among all of them. Answers can be words, values, phrases or sentences. The questions and their corresponding answers are of the elementary and middle school level extracted from science and history corpora. Two datasets are proposed: Chinese and English of five domains which are: *Biology*, *Chemistry*, *Physics*, *Earth Science* and *Life Science*. The size of the English dataset is 2686 questions for the

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 203–207,
Taipei, Taiwan, November 27 – December 1, 2017. ©2017 AFNLP

training set, 669 questions for the development set and 2012 for the test set. Hereafter an example extracted from the *Earth Science* domain:

- Most tsunami are caused by:
  1. (A) Earthquakes
  2. (B) Meteorites
  3. (C) Volcanic eruptions
  4. (D) Collisions of ships at sea

3 System Description

In order to understand the principle behind MappSent approach, it is important to introduce the task for which it has been designed. We first introduce the Question-to-Question similarity task of SemEval shared task [1], then we present MappSent, the approach that has been designed for this task and finally, we present its adaptation to the multi-choice question answering in examinations task.

3.1 Question-to-Question Similarity Task

In community question answering, the question-to-question similarity task (Task3, SubtaskB in SemEval) consists of reranking 10 related questions according to their similarity with respect to a given original question. Candidates are labeled as *PerfectMatch*, *Relevant* or *Irrelevant*. The training and development datasets consist of 317 original questions and 3,170 related questions[2]. The test sets of 2016 and 2017 respectively consist of 70 original/700 related questions and 88 original/880 related questions. The official evaluation measure towards which all systems were evaluated is the mean average precision (MAP) using the 10 ranked related questions. The experimental results of MappSent (Hazem et al., 2017) have shown the best results on SemEval (2016/2017) question-to-question similarity task over state-of-art approaches.

3.2 MappSent Approach

MappSent approach aims at providing a better representation of pairs of similar sentences, paragraphs and more generally, pieces of texts of any length. A prior condition is to have a training data set of annotated pairs of sentences. The main idea is: given a set of similar sentences, the goal is to build a more discriminant and representative sentence embedding space. We first compute word embeddings of the entire corpus, then, each sentence is represented by an element-wise addition of its word embedding vectors. Finally, a mapping matrix is built using the SVD decomposition to project sentences in a new subspace. Similar sentences are moved closer thanks to a mapping matrix (Artetxe et al., 2016) learned from a training dataset containing annotated similar sentences. Basically, a set of similar sentence pairs is used as seed information to build the mapping matrix. The optimal mapping is computed by minimizing the Euclidean distance between the seed sentence pairs.

MappSent approach consists of the following steps:

1. We train a Skip-Gram [3] model using Gensim (Řehůřek and Sojka, 2010)[4] on a lemmatized training dataset.

2. Each training and test sentence is pre-processed. We remove stopwords and only keep nouns, verbs and adjectives while computing sentence embedding vectors and the mapping matrix. This step is not applied when learning word embeddings (cf.Step 1).

3. For each given pre-processed sentence, we build its embedding vector which is the element-wise addition of its words embedding vectors (Mikolov et al., 2013; Wieting et al., 2016; Arora et al., 2017). Unlike Arora et al. (2017) we do not use any weighting procedure while computing vectors embedding sum[5].

4. We build a mapping matrix where test sentences can be projected. We adapted Artetxe et al. (2016) approach in a monolingual scenario as follows:

   - To build the mapping matrix we need a mapping dictionary which contains similar sentence pairs.

[1] http://alt.qcri.org/semeval2017/task3/

[2] http://alt.qcri.org/semeval2016/task3/index.php?id=data-and-tools

[3] CBOW model had also been experienced but it turned out to give lower results while compared to the SkipGram model.

[4] To ensure the comparability of our experiments, we fixed the python hash function that is used to generate random initialization. By doing so, we are sure to obtain the same embeddings for a given configuration.

[5] We explored this direction without success.

- The mapping matrix is built by learning a linear transformation which minimizes the sum of squared Euclidean distances for the dictionary entries and using an orthogonality constraint to preserve the length normalization.
- While in the bilingual scenario, source words are projected in the target space by using the bilingual mapping matrix, in our case, original and related questions are both projected in a similar subspace using the monolingual sentence mapping matrix. This consists of our adaptation of the bilingual mapping.

5. Test sentences are projected in the new subspace thanks to the mapping matrix.

6. The cosine similarity is then used to measure the similarity between the projected test sentences.

3.3 MappSent Adaptation

Two ways of adapting MappSent to the question-answering task can be considered. The first approach illustrated in Figure 1 is to follow the same procedure as the question-to-question similarity task. This would consist on using annotated pairs of questions and there corresponding answers to build the mapping matrix. However, this approach may be counter-intuitive since answers are not similar to questions as opposed to the question-to-question similarity task where the strong hypothesis is the similarity between question pairs. Since the mapping matrix aims at representing sentences in a subspace based on a given criteria. One can assume that mapping pairs of questions and their correct answers in a new subspace as a plausible alternative. We denote this first adaptation by $MappSent_{QA}$.

The second approach illustrated in Figure 2 and denoted $MappSent_{QQAA}$ tends to keep the strong hypothesis of sentence pairs similarity. Hence, instead of building one mapping matrix to represent questions and answers, we built two mapping matrices, one that represent similar question pairs and the other one to represent similar answers pairs. Finally, for a given test question, we extract the most similar question in the training data. Then, we compute a Cosine similarity between its corresponding answer, and the four test candidates. We select as correct answer the test candidate with the highest similarity score.

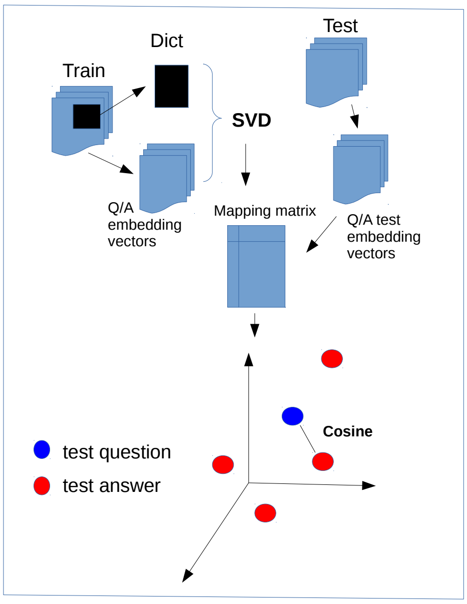

Figure 1: First adaptation: $MappSent_{QA}$

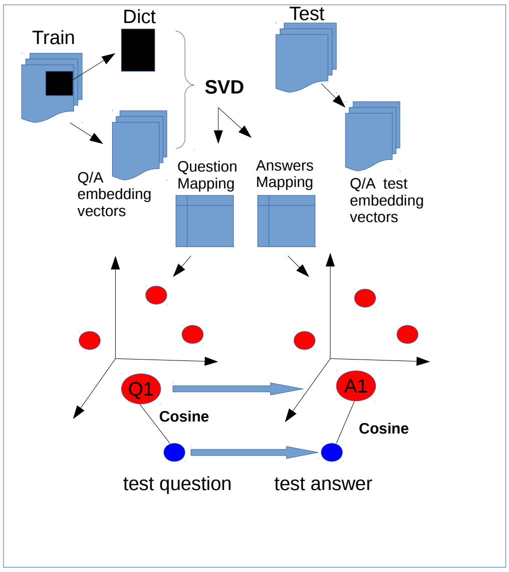

Figure 2: Second adaptation: $MappSent_{QQAA}$

3.4 Baseline

The baseline is a simple retrieval based approach which scores pairs of the question and each of its option as follows:

- concatenate a question with one of the candidate answers as a query
- use Lucene to search and extract relevant documents with the query

- score relevant documents by the similarity between the query and a given extracted document
- choose at most three highest scores to calculate the score of the pair of the question and the option
- output the pair with the highest score as the correct answer

4 Results

Method	Dev	Test
$Baseline$	29.45	-
$MappSent_{QA}$	33.3	29.5
$MappSent_{QQAA}$	**34.1**	**30.3**

Table 1: Results (Accuracy%) on IJCNLP 2017 shared task

Table 1 presents the results on the multi-choice question answering in examination task. We compare the two adaptations of MappSent approach that is: $MappSent_{QA}$ and $MappSent_{QQAA}$ to the baseline. We see that the two adaptations $MappSent_{QA}$ and $MappSent_{QQAA}$ outperform the baseline[6] on the development set. We see also that $MappSent_{QQAA}$ is slightly better than $MappSent_{QA}$ on both the development and the test sets. The best scores are 34.1% of accuracy on the development set and 30.1% of accuracy on the test set using $MappSent_{QQAA}$ approach.

5 Discussion

Several points have to be discussed regarding the obtained results of the two adaptations of MappSent. First, MappSent has been designed for question-to-question similarity. A direct application to question-answering pairs might be inappropriate. If the relations between similar question pairs are mainly lexical, semantics, reformulations, duplicates or near duplicates. Question-answering pairs are of a more complex relation nature, which can be pragmatic, rhetorical, elaboration, explanation, etc. This might explain the difficulties to capture these information and the mitigated results. Second, the task of multi-choice question answering exhibit specific particularities of the candidates. Answers can be words, values, phrases or sentences. In the case of words or values for instance, it is hard, if not impossible to represent the answer by an efficient embedding vector because of the lack of information conveyed by the candidates. Typically, our approach will always fail when answers are very short or contain only values. The third point that should be mentioned here is that MappSent didn't use any external data. This is of course an important drawback. The training data doesn't contain all the documents that might provide answers to the test questions. This is one of the clues that we let for future work. Another specificity of the multi-choice question is the fourth answer which can be of the form: ”All of these”, ”all of the above” or ”None of the above”, etc. Lets see the following example:

- In the process of cell division, the parent cell divides to form the:
  1. Continuation cells
  2. Next generation cells
  3. Daughter cells
  4. None of the above

We do not take into account this particular case in which the fourth answer: ”none of the above”, should be addressed in a special way. This is also a drawback that we need to deal with to improve our models.

6 Conclusion

In this paper, we have presented MappSent a novel and simple approach for textual similarity and proposed two adaptations for the task of multi-choice question answering. Our approaches allow to map sentences in a joint more representative sub-space. The experimental results have shown interesting results and lend support the idea that a mapping matrix is an appropriate method for textual representation and a promising approach for multi-choice question answering task. Furthermore, no attention has been given to external data and to the particularities of the multi-choice question answering data, drawbacks that we let for future work.

References

Sanjeev Arora, Liang Yingyu, and Ma Tengyu. 2017. A simple but tough to beat baseline for sentence

[6] We do not have yet the results of the baseline on the test set

embeddings. In *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, pages 1–11.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 2289–2294, Austin, TX, USA.

Yoshua Bengio, Rjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *JOURNAL OF MACHINE LEARNING RESEARCH*, 3:1137–1155.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning, ICML.*

Amir Hazem, Basma el amel Boussaha, and Nicolas Hernandez. 2017. Mappsent: a textual mapping approach for question-toquestion similarity. Recent Advances in Natural Language Processing, RANLP 2017, 2-8 September, 2017, Varna, Bulgaria.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *CoRR*, abs/1404.2188.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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. http://is.muni.cz/publication/884893/en.

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *Advances in Neural Information Processing Systems 24*.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2:207–218.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *CoRR*, abs/1503.00075.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. *Internationa Conference on Learning Representations, CoRR*, abs/1511.08198.