

# Improving Answer Selection with Analogy-Preserving Sentence Embeddings

Aïssatou Diallo<sup>1,2</sup>, Markus Zopf<sup>2</sup>, Johannes Fürnkranz<sup>2</sup>

<sup>1</sup> Research Training Group AIPHES

<sup>2</sup> Knowledge Engineering Group

Department of Computer Science, Technische Universität Darmstadt

Hochschulstraße 10, 64289 Darmstadt, Germany

{diallo@aiphes, mzopf@ke, juffi@ke}.tu-darmstadt.de

**Abstract.** Answer selection aims at identifying the correct answer for a given question from a set of potentially correct answers. Contrary to previous works, which typically focus on the semantic similarity between a question and its answer, our hypothesis is that question-answer pairs are often in analogical relation to each other. We propose a framework for learning dedicated sentence embeddings that preserve analogical properties in the semantic space. We evaluate the proposed method on benchmark datasets for answer selection and demonstrate that our sentence embeddings are competitive with similarity-based techniques.

**Keywords:** Analogical Reasoning · Embeddings · Answer Selection.

## 1 Introduction

Answer selection is the task of identifying the correct answer to a question from a pool of candidate answers. The standard methodology is to prefer answers that are semantically similar to the question. Often, this similarity is strengthened by bridging the lexical gap between the text pairs via learned semantic embeddings for words and sentences. The main drawback of this method is that question-answer (QA) pairs are modeled independently, and that the correspondence between different pairs is not considered in these embeddings. In fact, these methods focus only on each pair's entities relationship and are thus, limited to pairwise semantic structures. Instead, we argue in this paper that questions and their correct answers often form analogical relations. For example, the question "Who is the president of the United States?" and its answer are in the same relation to each other as the question "Who is the current chancellor of Germany?" and "Angela Merkel". Thus, for modeling these relations, we need to look at quadruples of textual items in the form of two question-answer pairs, and want to reinforce that they are in the same relation to each other.

---

An extended version of this work will appear at CoNLL-19.

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

We expect that using analogies to identify and transfer positive relationships between QA pairs will be a better approach for tackling the task of answer selection than simply looking at the similarity between individual questions and their answers.

We use sentence embeddings as the mechanism to assess the relationship between two sentences and aim to learn a latent representation in which their analogical relation is explicitly enforced in the latent space. Analogies are defined as relational similarities between two pairs of entities, such that the relation that holds between the entities of the first pair, also holds for the second pair. Loosely speaking, the quadruple of sentences is in *analogical proportion* if the difference between the first question and its answer is approximately the same as the difference between the second question and its answer. This formulation is especially valuable because analogies allow to put on relation pairs that are not directly linked. Consequently, in the vector space, analogous QA pairs will be oriented in the same direction, whereas dissimilar pairs will not correspond. The remainder of the paper is organized as follows: the next sections will shortly present related work on answer selection, the adopted approach for learning analogy-based embeddings and their evaluation on benchmark datasets for the task of answer selection.

## 2 Related Work

Answer selection is an important problem in natural language processing that has drawn a lot of attention in the research community [3]. Early works relied on computing a matching score between a question and its correct answer and were characterized by the heavy reliance on feature engineering for representing the QA pairs. Recently, deep learning methods have achieved excellent results in mitigating the difficulty of feature engineering. These methods are used to learn latent representations for questions and answers independently, and a matching function is applied to give the score of the two texts. Our work is also related to representation learning using deep neural networks. Although many studies confirmed that embeddings obtained from distributional similarity can be useful in a variety of different tasks, [4] showed that the semantic knowledge encoded by general-purpose similarity embeddings is limited and that enforcing the learnt representations to distinguish functional similarity from relatedness is beneficial. This work aims to preserve more far-reaching structures, namely analogies between pairs of entities.

## 3 Approach

The quadruple  $(a, b, c, d)$  is said to be in analogical proportion  $a : b :: c : d$  if  $a$  is related to  $b$  as  $c$  is related to  $d$ , i.e.,  $\mathcal{R}(a, b) \sim \mathcal{R}(c, d)$ . Let  $\mathcal{Q}$  and  $\mathcal{A}$  be the space of all questions and candidate answers. Enforcing the relational similarity between pairs of elements is equivalent to constraining the four elements to form a parallelogram in the Euclidean space. In an *analogical parallelogram*, there is

not only a relation  $\mathcal{R}$  holding between  $(a, b)$  and  $(c, d)$  respectively, but there must also hold a similar relation  $\mathcal{R}'$  between  $(a, c)$  and  $(b, d)$ . Given the aforesaid quadruple, when one of the four elements is unknown, an analogical proportion becomes an analogical equation which has the form  $a : b :: c : x$  where  $x$  represents an unknown element that is in analogical proportion to  $a, b, c$ . We aim at finding the element  $d_i$ , among  $n$  candidates, where the analogical proportion is as closely satisfied as possible. This allows us to re-frame the original problem of answer selection as a ranking problem, in which the goal is to select the candidate answer  $d$  with the highest likelihood to complete the analogy.

*Generating Quadruples* In order to create a dataset of quadruples to train the model, we adapt state-of-the-art datasets. Given a set of questions and their relative candidates answers, we construct the analogical quadruples in two steps. First, we divide all the questions into three different subsets of wh-word questions: "Who", "When" and "Where". We focus on these three types because their answer type falls in distinct and easily identifiable categories, namely answers of type "Location", "Person" and "Time". From these categories, we extract a variable number of prototypes, each of which is associated to a QA pair in the training set. Positive quadruples are obtained through the cartesian product between prototypes and the set of correct QA pairs of the same subset. Similarly, for negative training samples we associate a prototype, a question and a randomly selected wrong answer among its candidates. This approach will generate a set of hard examples to help improve the training.

*Quadruple Siamese Network* In order to tackle the aforementioned task, we propose to use a Siamese neural network, which are specifically designed to compute the similarity between two instances. The architecture has four sub-networks which share the learnt parameters. Every input sentence is encoded by bidirectional gated recurrent units (BiGRUs) with max pooling over the hidden states, and leads to a vector of dimension  $d$ .

In order to get the semantic relation between the pairs of input sentences, the network encodes the four  $d$ -dimensional embedding vectors, then merges them through a pairwise subtraction. For the energy of the model, we use the cosine similarity between the vector shifts of each pair of the quadruple as expressed in eq. (1). We use the contrastive loss [2] to perform the metric learning in eq. (2), which has two terms, one for the similar and another dissimilar samples. The similar instances are denoted by a label  $y = 1$  whereas the dissimilar pairs are represented by  $y = 0$ . Finally we sum the loss functions optimizing the two pairs of opposite sides of the parallelogram in eq. (3). The equations are the following:

$$E_{W_1} = |f_W(a - b) - f_W(c - d)| \quad \text{and} \quad E_{W_2} = |f_W(a - c) - f_W(b - d)| \quad (1)$$

$$\mathcal{L}_{W_i} = y (E_{W_i})^2 + (1 - y) \max((m - E_{W_i})^2, 0) \quad (2)$$

$$\mathcal{L} = \mathcal{L}_{W_1} + \mathcal{L}_{W_2} \quad (3)$$

This loss function forces analogous pairs to be mutually close and form an analogical parallelogram in the embedding space, while pushing dissimilar

transformations apart by a margin  $m$ , which we empirically choose to be 0.1. The parameters of the model are learned through a gradient based method that minimizes the L2-regularized loss. Further details about the implementation are given in section 4.

## 4 Experiments

We validate the proposed method on two datasets: WikiQA and TrecQA, two benchmark datasets for answer selection. We assess the performance of our method by measuring the Mean Average Precision (MAP) and the Mean Reciprocal Rank (MRR) for the generated quadruples in the test set. We initiate the embedding layer with FastText vectors. These weights are not updated during training. The dimension of the output of the sentence encoder is 300. For alleviating overfitting we apply a dropout rate of 0.6. The model is trained with Adam optimizer with a learning rate of 0.001 and weight decay rate of 0.01.

To support our claim that the learnt representations of our model encode the semantic of question answer pairs better than pre-trained sentence representation models, we choose four baselines commonly used to encode sentences: *Word2Vec and Glove* [6], *InferSent* [1] and *Sent2Vec* [5]. In order to minimize noise in the analogical inference procedure caused by comparing to multiple prototypes, we choose the quadruple with the highest analogical score.

Model	WikiQA	
	MAP	MRR
Glove	0.4640	0.4750
Word2Vec	0.4329	0.453
InferSent	0.3991	0.4037
Sent2Vec	0.4811	0.4866
<b>This work</b>	<b>0.6771</b>	<b>0.6841</b>

Table 1: Evaluation on quadruples.

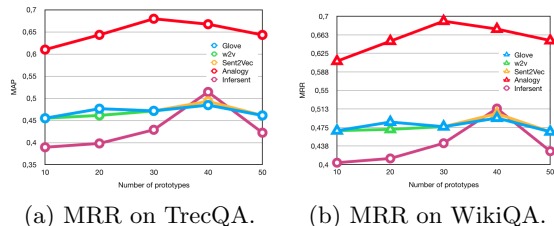


Fig. 1: MAP and MRR for different numbers of prototypes.

We applied the described procedure to vectors obtained from our network as well as from the baseline representation methods. The results are shown in Table 1. We observe that averaging word embeddings such as Glove or Word2Vec performs better than the dedicated sentence representations in the WikiQA dataset. This might be due to the fact that word embeddings have shown to encode some analogical properties. On the other hand, sentence embeddings have been trained with a particular learning objective, for example, InferSent has been trained for the task of claim entailment with a classification objective and might not be suitable for representing relations between pairs of sentences. Nevertheless, ranking by the cosine similarity of the difference vectors do not lead to acceptable performances. This confirms our hypothesis that pre-trained

sentence representation do not preserve analogical properties. Similarly, we measure the influence of the number of prototypes on the performances. We vary the number of prototypes pair  $p \in \{10, 20, 30, 40, 50\}$  and measure the MRR for both datasets. The results are shown in Figures 1a and 1b. We can observe that the best performances are obtained for  $p = 30$ . The reason might be that a high number of prototypes brings more comparisons and increases the probability of spurious interactions between QA prototypes and QA pairs.

## 5 Conclusion

This work introduced a new approach to learn sentence representations for answer selection, which preserve structural similarities in the form of analogies. Analogies can be seen as a way of injecting reasoning ability, and we express this by requiring common dissimilarities implied by analogies to be reflected in the learned feature space. We showed that explicitly constraining structural analogies in the learnt embeddings leads to better results over the distance-only embeddings. We believe that it is worth-while to further explore the potential of analogical reasoning beyond their common use in word embeddings. The focus of this work has been on answer selection, but analogical reasoning can be useful in many other machine learning tasks such as machine translation or visual question answering. As a next step, we plan to explore other forms of analogies that involve modeling across domains.

## References

1. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
2. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)
3. Lai, T.M., Bui, T., Li, S.: A review on deep learning techniques applied to answer selection. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 2132–2144 (2018)
4. Levy, O., Remus, S., Biemann, C., Dagan, I.: Do supervised distributional methods really learn lexical inference relations? In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 970–976 (2015)
5. Pagliardini, M., Gupta, P., Jaggi, M.: Unsupervised learning of sentence embeddings using compositional n-gram features. arXiv preprint arXiv:1703.02507 (2017)
6. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014)