Word Sense Induction: similarity measure to induce word senses

Rustam Mussabayev, Olzhas Kozbagarov

Institute of Information and Computational Technologies, Pushkin str. 125, Almaty, Kazakhstan

**Corresponding author's e-mail: rmusab@gmail.com, o.kozbagarov@ipic.kz*

Abstract

This article presents the developed approach for the problem on word sense induction. The developed approach is based on the application of the developed similarity measure, which determines the level of relatedness of contexts in which target words are used. The application of this approach has been demonstrated in relation to the Russian language, a language with a rich morphology and free word order that complicates the problem under consideration. The approach was tested on the data sets that were used on the first shared task on word sense induction in at the Dialogue Conference 2018.

Keywords: natural language processing, word sense induction, polysemy, homonymy, similarity measure, word embeddings

1 Introduction

One of the tasks in the field of natural language processing, which is used in many applications and has a long history, is the task of word sense disambiguation, the task of determining in what sense a polysemous word is used in the given context. Of particular interest are unsupervised approaches, so called word sense induction approaches, since they do not involve the use of annotated corpora, dictionaries, and it is especially necessary for the Russian language, since for the Russian language there is no currently comprehensive lexical inventory like WordNet for English language.

The statement of this problem is formulated as follows: a set of polysemous words (including homonyms) are given in used various contexts and it is required to group contexts according to senses in which the word is used.

This paper presents the developed approach for the word sense induction problem. The approach uses words embeddings (which are obtained on the basis of models built on neural networks) of words that constitute contexts of target words, and then considers the context of target word as a bag of word embeddings. Next, the semantic similarity of contexts is determined through the developed similarity measure by applying it to contexts represented as bags of words embeddings, thereby finding the level of relatedness between all contexts of target word. Then, the developed clustering algorithm is applied, which, on the basis of the estimated levels of relatedness between contexts, groups them according to senses they define. The resulting clusters determines contexts in according to senses which target word convey and thereby highlight senses that the target words carry in the indicated contexts.

There are a lot of models that can be used to represent words as embeddings, for example, *word2vec*, *Glove*. In the given paper the contextualized embeddings model so called ELMO was used to assign to every word its embedding.

The approach was tested on three data sets in Russian that were used in the shared task of word sense induction at the Dialogue Conference (2018). The distinction of the data sets is that the first one is constructed using only polysemous words, the second only homonyms and the third one contains polysemous words and homonyms at the same time. The clustering evaluations was based on the Adjusted Rand Index metric.

2 Conclusion

The paper presented the developed approach on task of word sense induction. The developed approach showed results that were superior on 5% to the best results demonstrated on the data set that uses both polysemous words and homonyms. On other sets, the results were also among the best. So on data set with homonyms the developed approach could surpass the second best on 15% percent according to Adjusted Rand Index.

References

[1] Panchenko A, Lopukhina A, Ustalov D, Lopukhin K, Leontyev A, Arefyev N, Loukachevitch N 2018 RUSSE'2018: A Shared Task on Word Sense Induction and Disambiguation for the Russian Language *In Proceedings of the 24rd International Conference on Computational Linguistics and Intellectual Technologies (Dialogue'2018)*

[2] Peters M, Neumann M 2018 Deep contextualized word representations. *NAACL*