University *of Ljubljana*
Faculty *of Computer and Information Science*

Cross-lingual sense disambiguation

Jure Tič, Nejc Velikonja, and Sandra Vizlar

Abstract

This article describes our approach to Word in Context natural language processing problem. The existing work on the topic was researched and summarized. and in the end 2 models were identified as potential. Context2Vec and RoBERTa. Different methods for making the corpus were also discussed and some solutions were presented.

Keywords
detect context

Advisors: Slavko Žitnik

Introduction

In this paper we will attempt to tackle one of the most difficult areas of natural language processing, this being context recognition. Unlike in machine language, in natural language the same word can bear different meanings, depending on the context in which it was written or spoken. Such a word is called a homonym and can be very tricky for machines to recognize, as it often requires an advanced understanding of the language or even specific topic in which it was mentioned. Nevertheless recognizing context is very important because it enables us to make more informed decisions.

Our task is to prepare a Slovene corpus and then use a model to determine whether or not a chosen word is used in the same context in two different sentences. We will try to implement the Slovene version of the WiC SuperGLUE[1] task that will classify the contexts having the same meaning as true or false.

Related work

Multiple studies have allready been done on the topic of Word in context clasification, achiving different results.

One of the first studies on the subject was *WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations* [2]. This article describes the process of preparing and optimizing corpus for english variant of the WiC task. It then suggests different methods, such as Context2Vec, BERT and many others with which to solve the task of context recognition. Of the suggested methods BERT proved to be the most successfully.

The second study *TransWiC at SemEval-2021 Task 2: Transformer-based Multilingual and Cross-lingual Word-in-Context Disambiguation*[3] improved upon its predecessor. They employed Transformer models with which they also hoped to minimize dependency on specific language. They seceded in achiving around 90 percent classification accuracy

In the third study was **SkoltechNLP at SemEval-2021 Task 2: Generating Cross-Lingual Training Data for the Word-in-Context Task** [4]. Authors chose to use neural system based on the XLM-R. And conduct their experiments on multiple languages reaching the classification accuracy of 89 - 60 percent, depending on the language.

While not discussing the solutions to the context detection problem, SuperGLUE [1] is a framework for evaluation of Neural Language processing tasks with the aim of encouraging improvement in their solutions. It is made for English language and features 8 tasks in which improvements are yet to be made. One of these tasks is also Word in Context.

Proposed Methods

The main method we will try in our classification are Context2Vec. If we manage to complete this in time and the results will be satisfactory we may also try SloBERTa 2.0 [5] and compare the results.

0.1 Preparing Corpus

We prepared the Corpus with the help of Dictionary of Homonym. We looked up some of the homonyms that we thought were going to have most distinct meanings and than searched Gigafida [6] database. We obtained sentences containing the desired words with web scraping Gigafida database. We paired the

sentences containing the same word randomly and than manually determined weather their meaning is the same or not. Each sample in the corpus is represented by the class of the pair (whether the particular word has the same meaning in both sentences), the focus word and both sentences.

0.2 Corpus analysis

Currently we have built the corpus manually by annotating the pairs of sentences. We will expand this by trying to automatic annotation, which will have to be manually checked. The current corpus will be presented in the following figures. Currently the corpus consists of 630 examples.

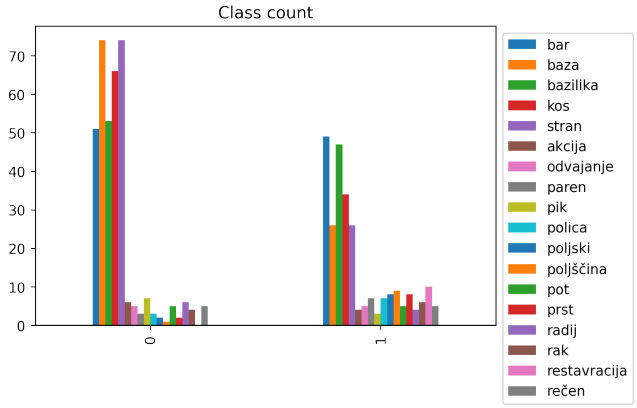

Figure 1. The count of each class for each word currently present in the corpus

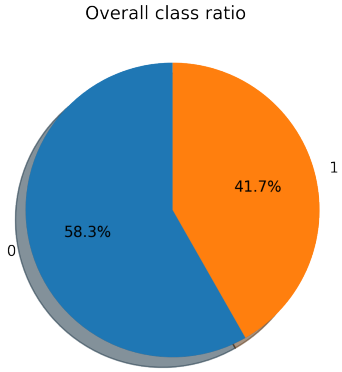

Figure 2. The proportion of each class in out corpus

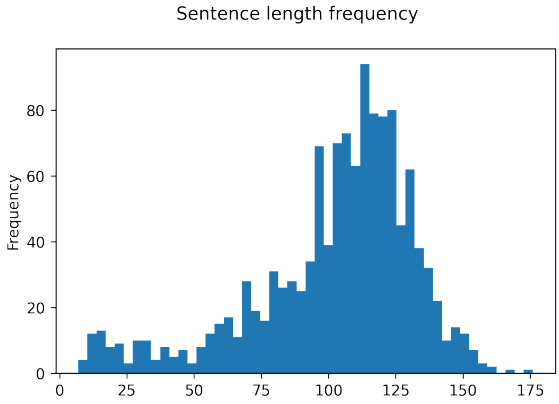

Figure 3. The histogram of sentence length for our corpus

We also compiled another automatically annotated corpus, which at the moment includes 10 words with 500 examples (sentence pairs) each. This corpus is automatically annotated with KMeans clustering based on TFIDF scores.

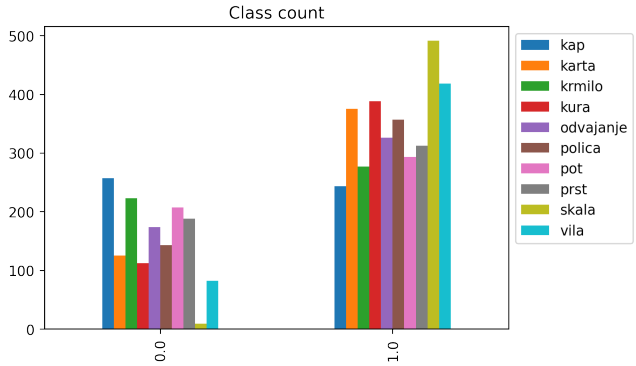

Figure 4. The count of each class for each word currently present in the automatically annotated corpus

Results

Discussion

References

[1] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2019.

[2] Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Hansi Hettiarachchi and Tharindu Ranasinghe. Transwic at semeval-2021 task 2: Transformer-based multilingual

and cross-lingual word-in-context disambiguation. *CoRR*, abs/2104.04632, 2021.

[4] Anton Razzhigaev, Nikolay Arefyev, and Alexander Panchenko. SkoltechNLP at SemEval-2021 task 2: Generating cross-lingual training data for the word-in-context task. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 157–162, Online, August 2021. Association for Computational Linguistics.

[5] Matej Ulčar and Marko Robnik-Šikonja. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0, 2021. Slovenian language resource repository CLARIN.SI.

[6] Tomaž Erjavec Miha Grčar Peter Holozan Simon Šuster Nataša Logar Berginc, Simon Krek. Gigafida, 2021.