LDK 2023

Word in context task for the Slovene language

Timotej Knez and **Slavko Žitnik**
University of Ljubljana, Faculty of Computer and Information Science
{timotej.knez, slavko.zitnik}@fri.uni-lj.si

Abstract

In natural language, it is important to understand which meaning of a word is used based on its context. For this reason, a Word in Context task was designed where the model is presented with two sentences containing the same target word. The goal of the model is to recognise if the same sense of the word is used in both sentences. Over the years, many models for solving this task in the English language have been proposed. However, research on the Word-in-Context (WiC) task for the Slovene language has been limited by the lack of annotated data available in the Slovene language. In this paper, we construct a new Slovenian corpus for the WiC task that will enable future research in this area. The constructed corpus is comparable in size to the widely used WiC corpus in the SuperGLUE task. We also perform some tests using simple algorithms to validate the usability of the corpus.

1 Introduction

The Slovenian language, like many other languages, contains numerous words with multiple meanings. For instance, words like "gol" (naked/goal) and "klop" (tick/bench) can have different interpretations in various sentences. The ambiguity of such a word poses a challenge for many NLP tasks, as the models need to recognise the intended meaning based on the context. The goal of the Word-in-Context (WiC) task is to help the embedding models learn to recognise the context and differentiate between different meanings. The task is formulated such that a model receives a pair of sentences that both contain the same target word. The model needs to then recognise whether the same meaning of the two words is used in both sentences. The WiC task is also included in the SuperGLUE benchmark (Wang et al., 2019). Solving this task for the Slovene language is limited by the lack of appropriately annotated datasets containing Slovene sentences. As part of one of the possible student projects in the natural language processing course at the Faculty for Computer and Information science at the University of Ljubljana, the students annotated a small number of sentences for the WiC task and used them to try and solve the task for the Slovene language. In this paper, we combined their manually annotated sentences into a single dataset that can be used for the Slovene Word in Context task. We also included a larger number of automatically annotated examples to help train models that might require a larger amount of data. We also used a number of simple models for the WiC task to demonstrate the usability of the constructed corpus. We compared the results achieved on our dataset to the results achieved with the same algorithms on the English dataset. We found that our dataset is somewhat more challenging than the English one due to some words with multiple similar meanings. The dataset is published in the Clarin.si repository[1].

2 Related work

The goal of this paper is to enable the Word-in-Context (WiC) task in the Slovene language. The Word-in-Context task was described by Wang et al. (Wang et al., 2019) as part of the SuperGLUE benchmark. The task is defined as a binary classification, where the model is presented with two sentences that contain a common homonym. The goal is for the model to recognise whether the same meaning of the target word is used in both sentences.

2.1 Datasets for the Word-in-Context task

The most commonly used dataset for the Word-in-Context task is the WiC dataset (Pilehvar and Camacho-Collados, 2018), provided by the SuperGLUE benchmark. The dataset contains around 7500 sentence pairs compiled from WordNet, Wiktionary, and VerbNet. Recently a larger version

[1] http://hdl.handle.net/11356/1781

of the dataset was published under the name XL-WiC (Raganato et al., 2020) which in addition to the English sentence pairs from (Pilehvar and Camacho-Collados, 2018), contains sentences from multiple other languages. The dataset contains training sets in three additional languages (German, French, and Italian) and validation and test sets in 12 additional languages. The goal of the dataset is to support cross-lingual inference. The sentence pairs were extracted from wiktionary and the multilingual WordNet.

A related dataset for the Finnish, Croatian, and Slovene languages was presented by Wand et al. (Armendariz et al., 2019). The dataset is designed for the word similarity in context task where we need to predict the semantic similarity between two different words based on the context presented in two sentences. They constructed the dataset by manually annotating sentence pairs based on how similar the two words are.

2.2 Models for solving the WiC task

El-Gedawy (El-Gedawy, 2013) presented a method for determining the meaning of Arabic words based on their context. They construct a dataset from WordNet. To improve the results, they provide the model with the most frequent words that appear when searching the sentence on Google and Bing search engines. This way the model gets information about the context of the sentence. The classification is performed by computing similarity between observed terms and terms from all word senses. The model manages to achieve an f-score of 80%. They also recognise, that removing stop words increases model performance.

Another approach for the task was proposed by Pal et al. (Pal et al., 2013). They use a model combining the bag-of-words approach with a Modified Lesk algorithm. The bag-of-words model is used to find the meaning of the ambiguous word. They construct a bag for each sense of the word. The sentence with removed stop words is compared to the words in each of the bags to determine the most likely sense. The Modified Lesk algorithm is used to detect word sense without supervision. While on its own it does not provide good performance, it improves the results when used in combination with the bag-of-words approach. The bag-of-words alone achieves 66% F-score, while the addition of the Modified Lesk algorithm improves the F-score to 85%.

Another interesting approach for word sense disambiguation was presented by Chaplot and Salakhutdinov (Chaplot and Salakhutdinov, 2018). The approach detects the topics that appear in the entire text instead of relying solely on the sentence the word is located in. The senses of the words are predicted based on the topics that appear in the document. The topic detection is performed using the Latent Dirichlet Allocation (LDA).

3 Dataset construction

In this section, we present an explanation of our pipeline for constructing a WiC corpus. The corpus was compiled from six student projects, where each group prepared a small dataset for the word-in-context task. As all groups followed a similar methodology, we present the combined process. An overview of the pipeline is depicted in Figure 1.

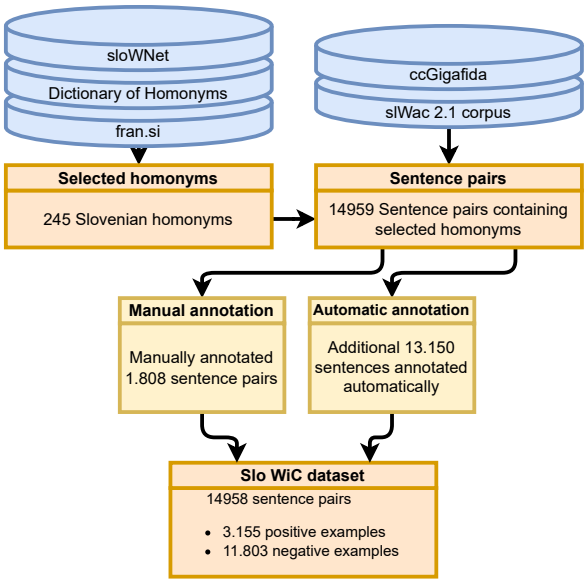

Figure 1: An illustration of the pipeline for constructing the Slovenian WiC dataset.

The first step in constructing the Slovenian corpus for the Word in Context task is to gather a list of homonyms to be included in our corpus. We gathered the homonyms from the Slovene dictionary of Homonyms (Bálint, 1997), Slovene wordnet (Fišer, 2015), and by scraping the Slovene dictionary website *Fran.si*. Once we had the interesting words to include in the dataset, we collected the sentences where the selected homonyms appear in different contexts. The sentences were gathered by searching the ccGigafida corpus (Logar et al., 2013) for the selected homonyms. The ccGigafida is a large corpus of Slovenian text. One group gathered the

sentences from the slWaC-Slovene web corpus.

Once the sentences were gathered, we need to annotate them to be used as training examples. We used a combination of manual annotations and automatic annotations computed by multiple machine-learning models. The process of manual annotation was performed in a few different ways by different groups. Most of the corpus was annotated by first constructing sentence pairs and manually annotating them with a label that shows whether the target word is used in the same sense in both sentences. On the other hand, one group first annotated a number of sentences with the senses of the target word. After that, they formed pairs of annotated sentences to get combinations of the different senses.

In addition to the manually annotated sentence pairs we also prepared some automatically labelled sentence pairs. The labels for these pairs were computed by clustering the sentences based on multiple algorithms. We used contextualized word embeddings computed by the BERT model, sentence embeddings based on Glove and Word2Vec embeddings, and bags of words. The labels were then determined by observing the similarity between both sentences. This approach produces some errors in annotations. To combat that we discarded the sentence pairs where the similarity scores were close to the threshold and only kept the pairs with very high and very low similarity. We manually analyzed a random sample of the automatically annotated corpus and found that the relations have 76% accuracy.

3.1 Dataset structure

For using the constructed corpus, it is important to understand its structure and parameters. As described in Section 3, a part of the corpus was annotated manually, while the other part contains automatically generated annotations. Altogether there are 7855 sentence pairs annotated manually and 7103 sentence pairs with only automatic annotations. Another important piece of information is how many times the same sentence can occur in the dataset. A large majority of the sentences appear in no more than four different sentence pairs. While some of the sentences appear in multiple sentence pairs, a large majority of the sentences appear in only a single sentence pair. 74% of all sentence pairs in the dataset contains only sentences that do not appear in any other sentence pair.

For training, it is important that the dataset is not too imbalanced. To check that, we analyzed the distribution of both classes. The manually labelled portion of our dataset contains 1200 sentence pairs (66.4%) that have the same meaning in both sentences and 608 sentence pairs (33.6%) with different meanings. In the entire corpus, there are 11803 sentence pairs (78.9%) with the same meaning and 3155 sentence pairs (21.1%) with different meanings. We found that the classes are a bit imbalanced; however, we believe that the level of imbalance is acceptable. Because of the imbalance we used the AUC measure in our tests instead of the classification accuracy.

Table 1: Comparison of the size of our word in context dataset and the English WiC dataset.

Corpus	**Sentence pairs**
English WiC - Train	5428
English WiC - Val	638
English WiC - Test	1400
English WiC - Sum	**7466**
Slo WiC - Manual	1808
Slo WiC - Automatic	13150
Slo WiC - Sum	14959

3.2 Comparison to the WiC dataset

We compare our Slovenian word in context dataset to the widely used English WiC dataset (Pilehvar and Camacho-Collados, 2018). When taking into account all of the annotated sentence pairs in our dataset including the automatically labelled examples, our dataset contains 14959 sentence pairs, which is larger than the English WiC dataset which contains 7466 sentence pairs. However, the automatically labelled examples might not be useful in all use cases as they might contain errors. Because of that the more appropriate comparison would be to observe the manually annotated part of our dataset, which contains 1808 sentence pairs. We present the size comparison of both corpora in Table 1.

Another important metric is the number of homonyms captured in the dataset. The English WiC dataset compares 2345 unique words. While our Slovenian WiC only contains 245 unique homonyms. That is because we include a larger number of sentence pairs for each homonym. We present the number of unique homonyms contained in each part of the two datasets in Table 2.

Table 2: Comparison of the number of homonyms contained in our word in context dataset and the English WiC dataset.

Corpus	Homonyms
English WiC - Train	1265
English WiC - Val	599
English WiC - Test	1184
English WiC - Combined	2345
Slo WiC - Manual	228
Slo WiC - Automatic	240
Slo WiC - Combined	245

4 Word in context models

Once we constructed the Slovenian Word in Context dataset, we can use it to train a WiC model. We constructed several models for solving the Word in Context task.

4.1 Clustering based prediction

The main approach that we used is based on clustering the sentences together. The goal is that we compute a contextual embedding of both sentences that captures the context in which the words are used. After that, we compute the distance between the embeddings to determine if the contexts are similar. For that, we need to determine a threshold similarity value based on the training data. Here we are working under the assumption that when a homonym is used in the same context, its sense will also be the same and vice versa.

For computing the distance between sentence embeddings we used cosine similarity. We tested multiple different methods for generating sentence embeddings to represent the context of each target word. A potential problem with this approach is that the assumption that when the word is used in different contexts its meaning will also be different might not always hold. On the other hand, the approach has a large advantage in that it is unsupervised and only requires training data to determine the similarity threshold.

4.2 Bag-of-words algorithm

To establish a baseline for our results, we utilized the Bag-of-words technique as a basic and straightforward approach. To implement this method, we utilized sentences that had already been stripped of stopwords. We kept track of the words that were in close proximity to the target word and represented them as a single large vector. By tallying the number of times these words appeared, we generated a vector for each sentence. To determine whether a target word was used similarly in two given sentences, we measured the cosine similarity between their respective vectors and applied a thresholding technique. Our Bag-of-words method takes the following parameters into account:

- Window size: This determines how many adjacent words around the target word will be used as context.
- Cosine distance threshold: If the cosine similarity between two vectors exceeds this predetermined threshold, the pair is deemed to have the same context.

4.3 The Simplified Lesk algorithm

We experimented with a simplified version of the Lesk algorithm as another method for solving the WiC task. For this algorithm, we used the sentences from our dataset with the stopwords removed. The Simplified Lesk algorithm works by comparing the sentence with a sample sentence with a known meaning. For the sample sentences we used the entire Dictionary of Standard Slovene Language (SSKJ) from a Github repository[2]. We computed the overlap between the lemma forms of the words that occurred in the sentences and the words in dictionary glosses of different meanings. During the preprocessing step, we stored the glosses in a dictionary based on the target words for efficient search. We also precomputed the lemmas of the words in glosses so that we could compare them with our sentence pairs. We used the CLASSLA pipeline (Ljubešić and Dobrovoljc, 2019) for extracting the lemma forms of all words used by this algorithm. This approach is especially interesting as it determines the meaning of the target word in each sentence and not only if the words in both sentences have the same meaning.

4.4 Pretrained language models

In recent years, many natural language tasks rely on using large pretraind language models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) for computing token embeddings. The main advantage of such models compared to using precomputed token embeddings is that they produce contextualized token embeddings which capture not only the information about

[2] https://github.com/van123helsing/SSKJ

the token but also about its context. Because of this, such models are very useful for differentiating between different meanings of the same word. Once we had the embeddings, we compared them using cosine distance to determine if the words are likely used in the same context. The architecture of the approach is shown in Figure 2

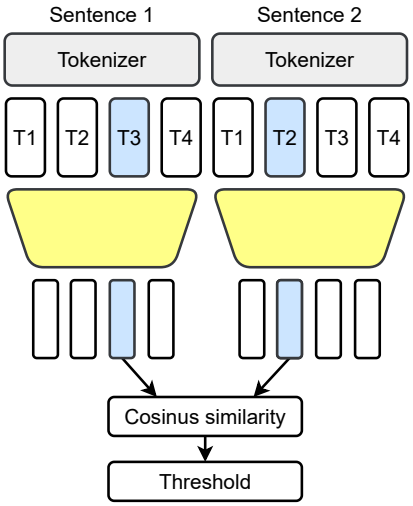

Figure 2: Architecture of the clustering model based on a pretrained language model.

In our tests, we used multiple pretrained BERT networks that are able to analyze Slovene text to produce contextualized embeddings of the target word in each sentence. The first network that we used is the Multilingual BERT model that was trained on 102 languages including Slovene. The second pretrained language model that we used is the CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020) which was trained on Croatian, Slovene, and English languages. The final pretrained language model that we used is the SloBERTa (Ulčar and Robnik-Šikonja, 2021) which was trained on just Slovene text. The multilingual models here have the advantage of being trained on a larger amount of data; however, that also means that they might not be well fitted to the Slovene language. On the other hand, SloBERTa is well fitted to the Slovene language but was trained on a much smaller corpus.

5 Results

We tested the presented methods for detecting if the same sense of the target word is used in both sentences in a sentence pair. The methods based on cosine similarity provide a score that needs to be compared with a threshold value. Instead of determining a single threshold value, we decided to evaluate the algorithms by observing the area under the ROC curve as we change the threshold. The curves are shown in Figure 3. The simplified Lesk algorithm provides classifications instead of some likelihood scores that could be compared to the threshold. Because of that, its performance is denoted by an x in Figure 3. We computed the AUC scores of all algorithms and presented them in Table 3. We also tested the best-performing algorithms on the English dataset (Pilehvar and Camacho-Collados, 2018) for comparison.

Table 3: The area under the curve scores of all tested algorithms. We also include scores on the English dataset for the best-performing multilingual approaches for comparison.

Embedding method	Slo AUC	Eng AUC
Random baseline	50%	50%
Bag-of-words	56.1%	
CroSloEngual BERT	68.9%	71.7%
Multilingual BERT	65.6%	68.5%
SloBERTa	55.5%	
Simplified Lex	58.7%	

All of the models were tested on the manually annotated part of the Slovene WiC corpus. We did not use the automatically generated part of the corpus as the proposed models do not benefit from a larger dataset and we wanted the results to be as accurate as possible.

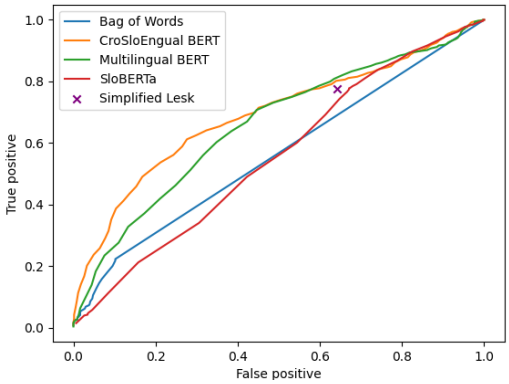

Figure 3: ROC curves of the predictions by the tested algorithms.

We found that the Simplified Lex algorithm achieved similar results as cosine similarity using the BERT embeddings. As expected the bag-of-words algorithm achieved worse results. The results are not directly comparable to the results achieved by previous research as the models were

tested on a different dataset.

5.1 Discussion

When using the clustering models, we are assuming that when two contexts of a word are different, the meaning of the word will be different as well. This assumption is somewhat problematic as the same meaning of a word might be used in multiple different contexts. In this case, the distance between the sentence embeddings might be large even though the meaning of the target word is the same. This aspect is improved by the Lesk algorithm, which compares the sentence to all known meanings of the word, which means that even if the two sentences fall under different clusters, they might get assigned the same meaning.

We also compared the scores achieved on the Slovene dataset to the ones achieved by the same algorithms on the English dataset. We found that the algorithms perform better when used on English data. The reason for this is likely that we included a number of words that have multiple very similar meanings that might be used in the same context. We believe that difficult words like this make the dataset better as they teach the model to differentiate between similar meanings.

Acknowledgment

The work presented in this paper was done as part of student projects in the Natural language processing course at the Faculty of computer and information science at the University of Ljubljana. We combined work done by the following students: Anže Luzar, Anže Tomažin, Blaž Beličič, Marko Ivanovski, Matej Miočić, Matej Kalc, Zala Erič, Miha Debenjak, Denis Derenda Cizel, David Miškić, Kim Ana Badovinac, Sabina Matjašič, Nejc Velikonja, Jure Tič, and Sandra Vizlar.

A part of this research was financially supported by the Slovenian Research Agency in the young researchers grant.

References

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, Marko Robnik-Šikonja, Mark Granroth-Wilding, and Kristiina Vaik. 2019. Cosimlex: A resource for evaluating graded word similarity in context. *arXiv preprint arXiv:1912.05320*.

Júlia Bálint. 1997. *Slovar slovenskih homonimov: na podlagi gesel Slovarja slovenskega knjižnega jezika*. Znanstveni Institut Filozofske Fakultete.

Devendra Singh Chaplot and Ruslan Salakhutdinov. 2018. Knowledge-based word sense disambiguation using topic models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Madeeh Nayer El-Gedawy. 2013. Using fuzzifiers to solve word sense ambiguation in arabic language. *International Journal of Computer Applications*, 79(2).

Darja Fišer. 2015. Semantic lexicon of slovene sloWNet 3.1. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of slovenian, croatian and serbian. In *Proceedings of the 7th workshop on balto-slavic natural language processing*, pages 29–34.

Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar, and Peter Holozan. 2013. Written corpus ccGigafida 1.0. Slovenian language resource repository CLARIN.SI.

Alok Ranjan Pal, Anirban Kundu, Abhay Singh, Raj Shekhar, Kunal Sinha, et al. 2013. An approach to word sense disambiguation combining modified lesk and bag-of-words. *Comput. Sci. Inform. Technol*, 3:517–524.

Mohammad Taher Pilehvar and José Camacho-Collados. 2018. Wic: 10,000 example pairs for evaluating context-sensitive representations. *arXiv preprint arXiv:1808.09121*, 6.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. The Association for Computational Linguistics.

Matej Ulčar and Marko Robnik-Šikonja. 2020. CroSloEngual BERT 1.1. Slovenian language resource repository CLARIN.SI.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.