Slovene Word Sense Disambiguation using Transfer Learning

Zoran Fijavž
University of Ljubljana, Faculty of Education
Slovenia
zoran.fijavzz@gmail.com

Marko Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science
Slovenia
marko.robnik@fri.uni-lj.si

ABSTRACT

Word sense disambiguation is an important task in natural language processing and computational linguistics with several practical applications, such as machine translation and speech synthesis. While the bulk of research efforts are targeted to English, some multilingual resources which include Slovenian have emerged recently. We utilized the Elexis-WSD dataset and a multilingual large language model to train models for word sense disambiguation in Slovenian, using sentence pairs with matching lemmas and matching or different word senses. The best model achieved an F_1 score of 81.6 on a Slovenian test set, although the latter had a restricted vocabulary due to filtering and is not comparable other testing frameworks. The exhaustive generation of sentence pairs for given lemmas and senses did not improve model performance and reduced the performance in out-of-vocabulary testing. Training on a mixed English-Slovene dataset maintained high test set as well as out-of-vocabulary results.

KEYWORDS

word sense disambiguation, transfer learning, multilingual transformer

1 INTRODUCTION

Word sense disambiguation (WSD) aims to identify the correct word sense used in a particular context. It is a long-standing problem in the field of computational linguistics and is important for downstream applications, such as machine translation, information retrieval, text mining, and speech synthesis. Recent WSD approaches use pre-trained large language models such as BERT [3], fine-tuning them on annotated data. As with most supervised machine learning approaches, there is a bottleneck on high-quality training data acquisition. The problem is severe, as standard WSD approaches treat each word sense as a separate target label. A partial solution is to use multilingual pretrained models that can leverage several WSD datasets.

In this paper, we demonstrate a methodology for cross-lingual transfer learning for WSD in Slovene that does not require compatible sense inventories in different languages. The proposed approach also works on out-of-vocabulary data.

After outlining related works in Section 2, we describe WSD models we developed for Slovene in Section 3, and their evaluation in Section 4. In Section 5, we provide an interdisciplinary critique of the current approaches to WSD that may be informative for future research. Section 6 presents the conclusions and ideas for further work.

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

2 RELATED WORK

One of the first WSD algorithms was Lesk [11] and its various extensions that are based on the word overlap between pre-defined sense definitions and target sentences. Conceptually, modern approaches to WSD remain strikingly similar, with advances stemming mostly from increasingly complex word representations (e.g. contextual word embeddings) and expansive lexicographical resources (e.g. a gloss list for word senses in SemCor). Recent approaches use supervised learning directly on word sense annotations [5], enrich sense definitions with various lexicographical resources [7, 19] and include lexical databases as graph data in conjunction with contextual word embeddings [2].

Until recently, the development of contemporary WSD models for Slovenian has been hindered by a lack of available datasets. That was partly addressed by the inclusion of Slovenian in the multilingual Elexis-WSD and XL-WSD datasets [12, 16]. Models trained on the latter obtained an F_1 score of 68.36% for Slovene WSD, which is significantly lower than state-of-the-art English models scoring 80% or above (although differing test frameworks preclude direct comparisons).

3 METHODOLOGY

In this section we describe the training procedure, data preparation and testing framework used to develop and test the Slovenian WSD models.

3.1 Training Task and Setup

We operationalized WSD as a sentence-pair binary classification task that distinguishes between sentence pairs with an identical or distinct word sense for a target lemma. Word senses were thus defined solely through annotated examples without the need for a secondary source of sense definitions (e.g. sense collocations, coarse semantic tags or glosses). Casting WSD as a binary classification task allowed us to combine Slovene and English datasets, as sentence pairs could be generated from different WSD datasets irrespective of sense inventory compatibility. Examples of the sentence pairs can be found in Table 1. The drawback of this approach was a significant data loss from filtering, as many lemmas did not have enough senses and use examples to generate sentence pairs.

For the base model, we used the pre-trained model CroSloEngual BERT [22] that can encode Slovenian, Croatian, and English texts. To reduce the training time and computational requirements, we used bottom layer freezing [10], gradient accumulation, and early stopping for non-converging models. Hyperparameter tuning was done on a 10% sample of the training data. We set the learning rate to 3e-5, gradient accumulation steps to 16, the batch size to 48, and the number of epochs to 2. Training a single model on 20% of all Slovenian sentence pairs required approximately 4 hours using a 16 GB NVidia GPU.

Table 1: Two Examples of the lemma *Cirkus* in the Pair Dataset and its English translation.

Lemma	Sentence 1	Sentence 2	Match
Cirkus	Družina na sliki s 'cirkusom' postuje po deželi.	Uprava 'cirkusa' ni odpovedala predstave.	Yes
Circus	*Family on the photo travels around the country with 'circus'.*	*The 'circus' management did not cancel the show.*	*Yes*
Cirkus	Uprava 'cirkusa' ni odpovedala predstave.	Zganjali so 'cirkus' okrog družinskih vrednot.	No
Circus	*The 'circus' management did not cancel the show.*	*They were making 'circus' around family values.*	*No*

Table 2: Number of Sentences, Lemmas and Word Senses in Datasets.

Datasets	Sentences (n)	Lemmas (n)	Word senses (n)
Original Sl.	202,240	5,604	11,069
Filtered Sl.	139,445	1,597	4,633
Full Sl. train	104,316	1,597	4,633
10% Sl. train	99,205	1,597	4,633
20% Sl. train	102,548	1,597	4,633
Validation	6,972	691	1,743
Test	28,157	1,597	4,633
10% En. train	27,028	2,852	9,683
20% En. train	27,123	2,852	9,683
20% mix train	126,233	4,437	14,316
OOV	3,006	25	50

3.2 Data Preparation

We used both Slovenian and English WSD datasets. The Slovenian data was obtained from the Slovenian section of the Elexis-WSD corpus [12] and the English data was drawn from SemCor to approximately match the size of the filtered Slovenian data.

Over 50% of the original Slovenian lemmas had a single sense tag. We removed multi-word and hyphenated senses and repeatedly filtered the datasets until there were at least two senses per lemma with at least four examples. The original dataset was thus heavily filtered from 202,240 sentences with 5,604 lemmas and 11,069 word sense tags to 139,445 sentences with 1,597 lemmas and 4,633 word sense tags. Punctuation was removed and target words were enclosed in apostrophes as a weak supervision signal [7].

The filtered Slovenian dataset was split into train, test and validation datasets. For the test dataset, we sampled two or eight sentences per word sense (depending on the total number of available sentences). The lower limit was needed to create sentence pairs and the upper limit was used to prevent frequent lemmas and senses from giving overly optimistic test scores. The validation dataset was created by sampling four sentences per word sense from lemmas with at least eight sentences, assuming frequent senses would be sufficient to detect over- and underfitting. The remainder of the data was kept for training. The Slovenian training and testing datasets contained the full coverage of included word Slovenian senses (4,633 distinct senses) and the validation dataset contained 1,743 senses. All Slovenian datasets included the full coverage of included lemmas (1,597). The Slovenian training dataset contained 104,316 unique sentences, the testing set 28,159 sentences and the validation dataset 6,972 sentences.

The filtered Slovene datasets were transformed into a dataset of sentence pairs by generating sentence combinations between sentences sharing a lemma. We limited the number of non-matching combinations generated to the number of possible matching combinations for each word sense. By storing infrequent sense pairs and downsampling frequent ones, we created two smaller Slovene sentence-pair datasets with the size of 10% and 20% of the original dataset.

The English dataset was created to complement the Slovenian one: we filtered out senses and lemmas that could not generate sentence pairs, filtered out infrequent lemmas, created a sentence-pair dataset and downsampled it to the size of the two smaller Slovenian datasets. The number of negative and positive pairs was roughly balanced for all pair datasets. Additionally, multiple smaller Slovene datasets [4, 13, 14, 17, 20, 21] were joined and filtered to create an out-of-vocabulary (OOV) dataset that included only lemmas absent from the main Slovenian dataset. The OOV dataset consisted of sentence pairs with matching or non-matching word senses for a target word. Table 2 summarizes the number of sentences, lemmas, and senses for each dataset.

In total, we trained 7 models that differed in the training data used: the entire Slovene dataset, the 10% Slovene dataset, the 20% Slovene dataset, the 10% English dataset, the 20% English dataset (with and without early stopping) and the mixed 20% dataset (a concatenation of the 10% Slovene and English datasets).

3.3 Evaluation Settings

Model performance was measured using the F_1 score and the Matthews correlation coefficient (MCC). The latter is a chi-square statistic computed from the confusion matrix of classification results. It served as an additional performance metric and enabled us to compare models without having to predict specific word sense tags (e.g., evaluate models on the OOV dataset with dissimilar lemmas and sense tags).

Two methods were used to predict the sense classes on the Slovenian test set. The first prediction method, called *the average sense probability heuristic* (ASP) used the test set structure with the models' binary classifier to determine the most likely sense. The target sentence was combined with all other test sentences sharing a lemma (except with itself) and a softmax value was obtained for each pair. The softmax values were averaged based on the sense tag of the non-target sentence and the sense with the highest average score was chosen as the sense prediction for the target sentence. The second prediction method used nearest neighbour matching between target sentence embeddings and *sense embeddings*. The latter were created by converting the entire Slovenian training and validation dataset into sentence embeddings [18] and averaging them by their word sense tags. The test sentences were likewise embedded and their sense label was predicted by selecting the sense embedding with the lowest cosine distance from the target sentence embedding.

The most frequent sense (MFS) heuristic as well as the sense embedding predictions from an untrained model were used as performance baselines. Lastly, several F_1 scores per model (micro-F_1, macro-F_1 and micro-F_1 by POS tags) were used as repeated

Table 3: F_1 Scores of Binary Classifier Predictions.

Model	Micro-F_1
MFS baseline	40.4
Full Sl.	81.0
10% Sl.	81.4
20% Sl.	80.5
10% En.	68.7
20% En.	46.9
20% En. (early stopping)	80.6
20% mix	**81.6**

Table 4: Binary Classifier MCC Test and OOV Scores.

Model	MCC test	MCC OOV
Full Sl.	**0.629**	0.273
10% Sl.	0.55	0.292
20% Sl.	0.578	0.284
10% En.	0.321	0.268
20% En.	0.004	0.273
20% En. (early stopping)	0.491	**0.353**
20% mix	0.578	0.326

Table 5: F_1 Scores of Nearest Neighbour Predictions.

Model	Micro-F_1
MFS baseline	40.4
Untrained model	21.7
Full Sl.	**72.8**
10% Sl.	50.9
20% Sl.	60.7
10% En.	53.2
20% En.	60.6
20% En. (early stopping)	28.7
20% mix	61.0

measures for model comparison using the Friedman test with the Nemenyi post-hoc test.

4 RESULTS

We evaluated model predictions with binary classifiers and with nearest neighbour matching to sense embeddings. Additionally, we used the Matthews correlation coefficient to evaluate the performance of binary classifiers and evaluate model performance on the out-of-vocabulary dataset.

4.1 Binary Classifier Sense Predictions

The baseline F_1 from the MFS heuristic was 40.4%. The difference between model predictions was statistically significant (χ^2_F = 36.12; df = 5; n = 8; p < 0.001) with the top three models differing significantly from the MFS baseline: the models, trained on the mixed 20% training data (F_1 = 81.6; p = 0.001), the 10% Slovene data (F_1 = 81.4; p = 0.026), the entire Slovene dataset (F_1 = 81; p = 0.004). Detailed results from predictions with binary classifiers can be found in Table 3. The statistical differences between binary classification models are presented in Figure 1.

4.2 Binary Classifier Correlation Metrics

As the testing set was transformable into sentence pairs, we used the binary classifiers directly on the test set and computed a MCC without predicting sense labels. We also applied the same procedure to test model performance on the OOV dataset.

The highest correlation between actual and predicted binary labels was achieved by the model, trained on the entire Slovenian dataset (MCC = 0.629) followed by models, trained on the 20% Slovene and 20% mixed datasets (MCC = 0.578; for both). The highest correlation between the actual and predicted labels on the OOV dataset was achieved by the model, trained on the 20% English dataset with early stopping (MCC = 0.353), followed by the 20% mixed dataset (MCC = 0.326). It should be noted that the former was a base model with minimal updates, as the training stopped after a single update at 200 out of 1916 total steps. Interestingly, ranking the models by the amount of included training data revealed a positive correlation between the number of included examples and the testing dataset MCC (r_s = 0.566; df = 5; p = 0.185) and a negative correlation between the number of included examples and OOV dataset MCC (r_s = -0.378; df = 5; p = 0.404), although neither association was statistically significant. Detailed results from MCC testing can be found in Table 4.

4.3 Sense Predictions with Nearest Neighbour Matching

For predictions with nearest neighbour matching between target sentence and sense embeddings, the baselines used were the MFS heuristic (F_1 = 40.4%) and the predictions from the untrained model (F_1 = 21.7%). The difference between model predictions was statistically significant (χ^2_F = 45.11; df = 5; n = 9; p < 0.001). The only model significantly different from the MFS predictions was trained on the entire Slovene dataset (F_1 = 72.8%; p = 0.003). Detailed results from predictions using nearest neighbour matching can be found in Table 5. The statistical differences between nearest neighbour predictions from different models are presented in Figure 2.

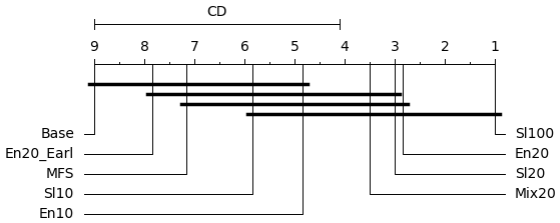

Figure 1: Critical Distance Diagram for Nearest Neighbour Results.

5 DISCUSSION ON INTERDISCIPLINARY ASPECTS

In this section, we offer a brief critique of the WSD task from the perspective of psycholinguistics, pragmatics and insights gained through model development, and suggest options for further research.

The datasets commonly used for WSD are not transparent in terms of the specific sense ambiguities they contain in spite of available typologies. Psycholinguistic literature has identified significant differences in human processing between homonymy

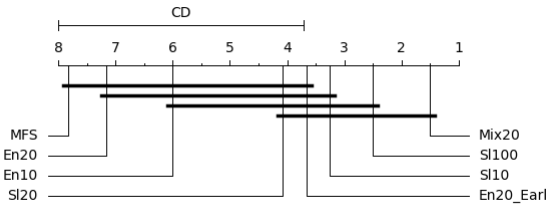

Figure 2: Critical Distance Diagram for Binary Classification Results.

and polysemy [8], as well as between various subtypes of the latter (e.g., metonymy and metaphors) [9]. As demonstrated by the use of the out-of-vocabulary test set, additional datasets, even if comparatively small, can provide important additional information on model performance. Incorporating a theoretically informed typology of polysemy or lexical ambiguity, future research could provide richer descriptions of word sense relations contained in widely used WSD datasets as well as develop specific tests for various types of polysemy. The latter could draw on datasets from psycholinguistic experiments, which commonly control for a plethora of variables, such as word and sense frequency. We also observed Elexis-WSD and SemCor contained a large number of single-sense lemmas, which would explain why F_1 scores from the MFS heuristic in related works are commonly relatively high.

Furthermore, while large language models have achieved state-of-the-art results in WSD, they do not fundamentally diverge from distributional semantics [6], which is but one account of possible disambiguation mechanisms. It is possible, for instance, to conceptualise word disambiguation as a pragmatic process whereby the common ground (shared knowledge) between speakers [1] scaffolds disambiguation and by which account speakers may introduce ambiguity on purpose to meet various communicative goals [15].

6 CONCLUSION

We developed several word sense disambiguation models for Slovenian text and achieved comparatively high performance, albeit on a limited selection of lemmas and word senses. We demonstrated that including small datasets to measure out-of-vocabulary performance yields important insights, as the models tended to generalize better with compacter training datasets.

The models presented in this paper would benefit from a review of Slovenian lexicographical sources and sense inventory compatibility between them. Replacing annotated sentences with sense definitions (e.g. collocation lists, coarse semantic tags, gloss definitions) would greatly increase the number of available training examples. Other large language models could also be used and a detailed hyperparameter optimization could be performed for each model individually.

The source code related to this paper and the datasets used are freely available[1].

Acknowledgments

The work was partially supported by the Slovenian Research and Innovation Agency (ARIS) core research programme P6-0411, and projects J6-2581 and J7-3159.

[1]https://github.com/zo-fi/slo_wsd_ZFMA

REFERENCES

[1] Keith Allan. 2013. What is Common Ground? In *Perspectives on Linguistic Pragmatics*. Perspectives in Pragmatics, Philosophy & Psychology. Alessandro Capone, Franco Lo Piparo, and Marco Carapezza, editors. Springer, Cham, 285–310. DOI: 10.1007/978-3-319-01014-4_11.

[2] Michele Bevilacqua and Roberto Navigli. 2020. Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2854–2864. DOI: 10.18653/v1/2020.acl-main.255.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. 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)*, 4171–4186. DOI: 10.18653/v1/N19-1423.

[4] Zala Erič, Miha Debenjak, and Denis Derenda Cizel. 2022. Cross-lingual sense disambiguation. GitHub repository. https://github.com/dextos658/Cross-lingual-sense-disambiguation.

[5] Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5297–5306. DOI: 10.18653/v1/D19-1533.

[6] Zellig S. Harris. 1954. Distributional Structure. *WORD*, 10, 2-3, 146–162. DOI: 10.1080/00437956.1954.11659520.

[7] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3509–3514. DOI: 10.18653/v1/D19-1355.

[8] Ekaterini Klepousniotou and Shari R. Baum. 2007. Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics*, 20, 1, 1–24. DOI: 10.1016/j.jneuroling.2006.02.001.

[9] Ekaterini Klepousniotou, G. Bruce Pike, Karsten Steinhauer, and Vincent Gracco. 2012. Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language*, 123, 1, 11–21. DOI: 10.1016/j.bandl.2012.06.007.

[10] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4365–4374.

[11] Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation* (SIGDOC '86), 24–26. ISBN: 978-0-89791-224-2. DOI: 10.1145/318723.318728.

[12] Federico Martelli et al. 2022. Parallel sense-annotated corpus ELEXIS-WSD 1.0. *https://elex.is/*. Retrieved Oct. 21, 2022 from https://www.clarin.si/repository/xmlui/handle/11356/1674.

[13] Matej Miočić, Marko Ivanovski, and Matej Kalc. 2022. NLP-tripleM. GitHub repository. https://github.com/KalcMatej99/NLP-tripleM.

[14] David Miškić, Kim Ana Badovinac, and Sabina Matjašič. 2022. cross-lingual-sense-disambiguation. GitHub repository. https://github.com/NLP-disambiguation/cross-lingual-sense-%20disambiguation.

[15] Brigitte Nerlich and David D. Clarke. 2001. Ambiguities we live by: towards a pragmatics of polysemy. *Journal of Pragmatics*, 33, 1, (Jan. 2001), 1–20. DOI: 10.1016/S0378-2166(99)00132-0.

[16] Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: an extra-large and cross-lingual evaluation framework for word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35, 15, 13648–13656. DOI: 10.1609/aaai.v35i15.17609.

[17] Erazem Pušnik, Rok Miklavčič, and Aljaž Šmaljcelj. 2022. nlp-project3. GitHub repository. https://github.com/RoKKim/nlp-project3.

[18] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. DOI: 10.18653/v1/D19-1410.

[19] Yang Song, Xin Cai Ong, Hwee Tou Ng, and Qian Lin. 2021. Improved Word Sense Disambiguation with Enhanced Sense Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 4311–4320. DOI: 10.18653/v1/2021.findings-emnlp.365.

[20] Jure Tič, Nejc Velikonja, and Sandra Vizlar. 2022. NLP. GitHub repository. https://github.com/JureTic/NLP.

[21] Andrej Tomažin. 2022. nlp-wic. GitHub repository. https://github.com/anzetomazin/nlp-wic.

[22] Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *Text, Speech, and Dialogue*, 104–111. DOI: 10.1007/978-3-030-58323-1_11.