Towards Slovene Word Sense Disambiguation with Transfer Learning

Zoran Fijavž
Faculty of Education, University of Ljubljana, Mirovni Inštitut

Prof. dr. Marko Robnik Šikonja
Faculty of information Faculty of Computer and Information Science

Ljubljana, 2023

Presentation structure

1. Intro & Motivation
2. WSD Model Development
3. Interdisciplinary Aspects
4. Future Work

Intro & Motivation

Motivation #1: Build WSD classifier for Slo

- Word sense disambiguation → downstream use for IR, MT, text mining, comp. lexicography
- Data acquisition bottleneck (e.g. OMSTI, SemCor)
- Transfer learning → Large number of labels (compared to NER, sentiment)

Motivation #2: Solved problem or problematic solution?

- Interdisciplinary aspects of WSD (psycholinguistics, pragmatics)

SloWSD model development

Learning task

- Sentence pair matching → same lemma, different sense
- Sense definition via examples of use (no external sense definitions)

Data sources:

- ElexisWSD: Slovenian part (Martelli et al. 2022)
- Selection from SemCor (Miller et al., 1994)
- Out-of-vocabulary dataset (various small Slo. WiC datasets)

Recap

Slo WSD development

- Classifier with limited scope (F_1 = 81,6; 4.633 lemmas)
- Forgetting through fine-tuning (not stat. significant)
- Multilingual benefits != just more data: density & diversity → generalization

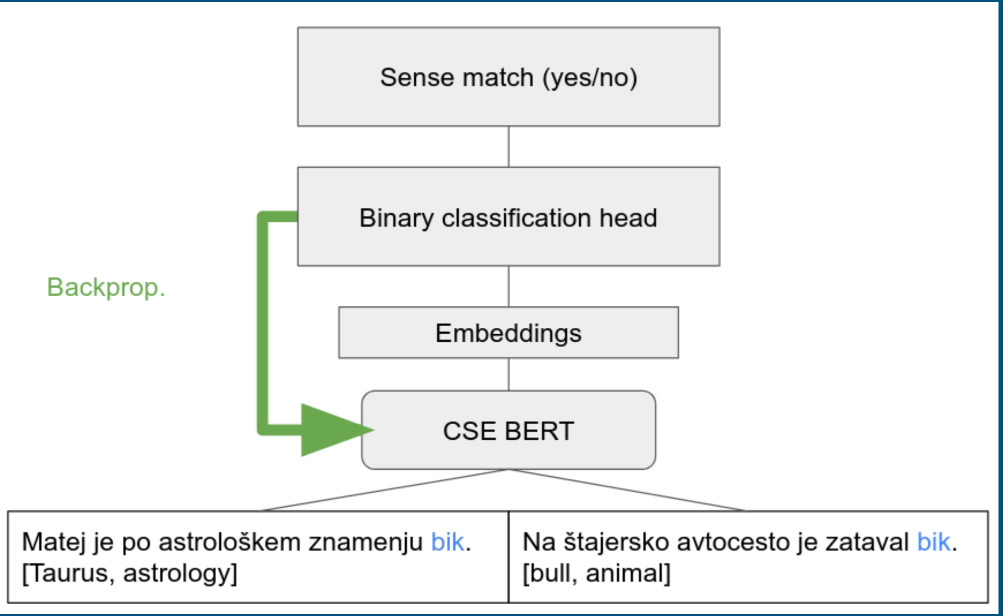
Interdisciplinary aspects

- Task definition matters
- Polysemy typologies not used in major WSD tasks
- Small data for multiple tests → richer descriptions

Data preparation

- Extensive filtering
  - MWU, punctuation removal & weak supervision (apostrophe)
  - 2 senses per lemma, enough examples for train-test split
  - Training, testing set = full coverage of sense labels
  - Dev set = selection from frequent senses
  - Test set = sampling with upper limit (very different distribution)
- Transformation into sentence pairs
  - Exhaustive combinations?
  - Downsampling (stratified by sense tag combinations)
  - Joining Slo. and Eng. datasets

Learning task visualization



Outline of supervised learning used

Learning task visualization

Lema	Pojavnica	Stavek	ID pomena	ID stavka
cirkus	cirkusom	Družina ki jo vidite na sliki pa s 'cirkusom' potuje po deželi	cirkus%0	6207
cirkus	cirkusu	Američana sta v teniškem 'cirkusu' dosegla skorajda vse kar se je doseči dalo	cirkus%1	6213
cirkus	cirkus	Ko so bili znani rezultati pregleda so zagnali cel 'cirkus'	cirkus%2	6234

Entries in the (basic) datasets

Learning task visualization

Stavek 1	Stavek 2	Oznaka ujemanja	ID pomenov
Družina ki jo vidite na sliki pa s 'cirkusom' potuje po deželi	Uprava 'cirkusa' ni odpovedala niti ene od naslednjih predstav	1	cirkus%0, cirkus%0
Uprava 'cirkusa' ni odpovedala niti ene od naslednjih predstav	Američana sta v teniškem 'cirkusu' dosegla skorajda vse kar se je doseči dalo	0	cirkus%0, cirkus%1
Američana sta v teniškem 'cirkusu' dosegla skorajda vse kar se je doseči dalo	Ko so bili znani rezultati pregleda so zagnali cel 'cirkus'	0	cirkus%0, cirkus%2

Entries in the dataset of sentence combinations

Base model, hyperparameters, settings

Base model: CroSloEngual BERT (Ulčar & Robnik-Šikonja, 2020)

Hyperparameter selection

- Learning rate, epoch num., batch size, gradient accumulation steps
- Probabilistic optimization on small Sl. training set)

Other configurations:

- Layer freezing (Merchant et al., 2020)
- Tokenizer max. Len. 180 (GPU issues)
- Early stopping

Testing framework

- Test scores: micro F_1 & Matthews Correlation Coefficient
    - MCC to evaluate sentence matching without labels → OOV testing
- Prediction #1: Binary classifier to sense labels
    - Test set structure with full coverage
    - Highest average softmax between related test sentences
- Prediction #2: Nearest Neighbour of target sentence
    - Sense embeddings from train & validation set
    - Testing the base model

Razvoj modelov za razdvoumljanje v slovenščini

7 models with different training data:

- Whole Slo. training set
- 10% & 20% Slo. training set
- 10% Eng. training set
- 20% Eng. training set (with and without early stopping)
- 20% mixed training set

Two baselines:

- Most frequent sense heuristic (train & val. set)
- Base CroSloEngual BERT (NN)

Model results

Out-of-vocabulary evaluation with Matthews correlation coefficient

- 20% Eng. set (early stopping; MCC = 0.353) → base model approximation
- 20% mixed set (MCC = 0.326)
- More sent pairs = worse OOV score (r_s = -0.378; df = 5; p = 0.404),

Prediction with NN of sense embeddings:

- Entire Slo. set (F_1 = 72.8)

Model results

Sense prediction with binary classifier:

- Best: 20% mixed set (F_1 = 81.6)
- Next: Whole Slo., 10% Slo., 20% Eng. set (base model approximation)

Binary predictions with MCC on sent. pairs:

- Best: Entire Slo. set (MCC = 0.629)
- Next: 20% mixed and 20% Slo. set (MCC = 0.578; both)

Interdisciplinary aspects of WSD

SOTA models approaching inter-annotator agreement

→ Solved problem or problematic solution?

1. What kind of multiple meanings?
- Existing typologies & differences in (human) processing:
  - Homonymy VS polysemy (Rodd et al., 2002; Klepousniotou in Baum, 2007)
  - Within polysemy: metaphors VS metonymy (Klepousniotou et al., 2012)
- Context & sense frequency as factors (MacDonald et al., 1994; Twilley et al., 1994)

Interdisciplinary aspects of WSD

2. Pragmatics of disambiguation

- Theory of mind (Apperly, 2012) VS distributional hypothesis (Harris, 1954)
- Infant studies: "I guess they want the new toy" (Tomasello & Haberl, 2003)
- Pragmatic reasoning, common ground, multimodality for disambiguation scaffolding

Interdisciplinary aspects of WSD

3. Dataset observations

- Disambiguation of single-sense lemmas
  - Prevalent in existing datasets
  - Multi-sense lemmas: Elexis-WSD 26.9%; SemCor 21% of lemmas
  - High MSF baseline + homonymy disambiguation = SOTA?
- Opaque descriptions of included polysemy/ambiguity
  - Dataset comparability (no control for sense typology)

Summary

Slo WSD development

- Classifier with limited scope (F_1 = 81,6; 4.633 lemma)
- Forgetting through fine-tuning (not stat. significant)
- Multilingual benefits != just more data: density & diversity ⟶ generalization

Interdisciplinary aspects

- Task definition matters
- Polysemy typologies not used in major WSD tasks
- Small data for multiple tests ⟶ richer descriptions

Future work

Other models, architectures, hyperparameters

Mapping sense inventories for Slo.:

- Sense definitions from external lex. sources (prevent data loss)

Specialized datasets for extensive WSD testing

- E.g. integration of existing typologies & datasets from psycholinguistics

References

Apperly, I. A. (2012). What is „theory of mind"? Concepts, cognitive processes and individual differences. Quarterly Journal of Experimental Psychology (2006), 65(5), 825–839. https://doi.org/10.1080/17470218.2012.676055

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

Harris, Z. S. (1954). Distributional Structure. WORD, 10(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Klepousniotou, E. & Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. Journal of Neurolinguistics, 20(1), 1–24. https://doi.org/10.1016/j.jneuroling.2006.02.001

Klepousniotou, E., Pike, G. B., Steinhauer, K. & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. Brain and Language, 123(1), 11–21. https://doi.org/10.1016/j.bandl.2012.06.007

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. https://doi.org/10.18653/v1/D19-1355

MacDonald, M. C., Pearlmutter, N. J. & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. Psychological Review, 101(4), 676–703. https://doi.org/10.1037/0033-295x.101.4.676

Merchant, A., Rahimtoroghi, E., Pavlick, E. & Tenney, I. (2020). What Happens To BERT Embeddings During Fine-tuning? Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 33–44. https://doi.org/10.18653/v1/2020.blackboxnlp-1.4

References

Miller, G. A., Chodorow, M., Landes, S., Leacock, C. & Thomas, R. G. (1994). Using a Semantic Concordance for Sense Identification. Human Language Technology.

Rodd, J., Gaskell, G. & Marslen-Wilson, W. (2002). Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access. Journal of Memory and Language, 46(2), 245–266. https://doi.org/10.1006/jmla.2001.2810

Tomasello, M. & Haberl, K. (2003). Understanding attention: 12- and 18-month-olds know what is new for other persons. Developmental Psychology, 39, 906–912. https://doi.org/10.1037/0012-1649.39.5.906

Twilley, L. C., Dixon, P., Taylor, D. & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. Memory & Cognition, 22(1), 111–126. https://doi.org/10.3758/BF03202766

Ulčar, M. & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. V: P. Sojka, I. Kopeček, K. Pala, in A. Horák (ur.), Text, Speech, and Dialogue (str. 104–111). https://doi.org/10.1007/978-3-030-58323-1_11