Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 402–408
https://doi.org/10.22364/bjmc.2022.10.3.13

Towards Word Sense Disambiguation for Latvian

Pēteris PAIKENS, Laura RITUMA, Lauma PRETKALNIŅA

University of Latvia Institute of Mathematics and Computer Science

peteris@ailab.lv, laura@ailab.lv, lauma@ailab.lv

Abstract. The goal of this paper is to describe the current situation on word sense disambiguation for Latvian, reviewing the available data and potential problems, and describing the exploration of word sense disambiguation methods using BERT contextual embeddings in order to apply them to Latvian language. Training is performed on a recently developed dataset of sense example sentences. The experiments of this paper demonstrate the feasibility of the approach by applying a mixture-of-experts approach of word sense disambiguation to the data, developing the first proof of concept WSD system for Latvian using state of art approaches. An evaluation of the WSD solution was performed on a selection of 18 highly ambiguous words, demonstrating reasonable performance.

Keywords: word sense disambiguation, Latvian, semantics

1 Introduction

Word Sense Disambiguation (WSD) is the task of associating words in context with their possible meanings contained in a pre-defined sense inventory. The goal of this project is to develop a word sense disambiguation system for Latvian based on the recently developed Latvian WordNet dataset (Paikens et al., 2022) that includes a substantial number of sentence examples matched to specific word senses.

In computational linguistics, tasks involving semantic analysis and natural language understanding (NLU) inevitably require treatment of word meaning and its ambiguity – either as a separate component explicitly performing word sense disambiguation (WSD), or by having information of possible word senses and their relations as additional data input to a NLP system that solves a particular task that implicitly involves WSD. Examples of NLP tasks that require WSD include semantic parsing, information extraction, information retrieval, abstractive text summarization and dialogue systems. WSD tools and sense-annotated corpora built with assistance of these tools are also of high value in lexicographic research and digital humanities.

The desire for a WSD system for Latvian has been relevant for a long time, but as of now no WSD solutions were available as only recently an appropriate sense inventory and data for training and evaluation became available. In this paper we describe the application of current state of art research on English WSD to develop such a solution for Latvian.

2 Related work

For Latvian language, there has been no published research of general purpose automatic word sense disambiguation. The closest relevant work is the analysis on word sense disambiguation linguistic principles used in preparing this dataset (Lokmane et al., 2021) and earlier work on word sense disambiguation in the very restricted domain of controlled natural language using logic reasoning (Bārzdiņš et al., 2007).

On the other hand, for other languages word sense disambiguation has been a widely researched topic. Most of the advanced research has been performed on English, with later application of these approaches to other languages. The primary restriction on applying these methods to Latvian has been the lack of an appropriate word sense inventory and data to develop and evaluate such systems, but such data has been recently made available (Paikens et al., 2022). However, it is plausible that adapting these methods to Latvian may require research and modifications, as differences in linguistic properties such as morphological variation and less strict word order often require changes in NLP methods used (Bender, 2011).

The most commonly used approach for current state of art WSD systems for English rely on training many lemma specific classifiers ("word experts") for disambiguating senses of that lemma. This approach was successfully used both before widespread application of contextual word embeddings (Iacobacci et al., 2016) and - with significantly improved results - after applying the improved embeddings from BERT (Devlin et al., 2019) and related models (Hadiwinoto et al., 2019; Vial et al., 2019).

An alternative approach applies pretrained language models in a more direct manner, adapting the task as sentence pair classification, which is one of the main tasks for the pretrained BERT models. It can be done for the sentence context paired with each of candidate glosses (Huang et al., 2019; Blevins and Zettlemoyer, 2020) or, interestingly, as concatenating all the glosses for the target word in a single 'sentence' and attempting span extraction to determine most appropriate choice (Barba et al., 2021). For multilingual approaches, zero-shot learning from multilingual embeddings achieves competitive results (Pasini et al., 2021).

While these approaches are technically different, they achieve similar performance on English datasets. Complex models that integrate many different types of data achieve an accuracy improvement of around 2 percentage points (Song et al., 2021), but in this early stage of research these accuracy differences are less significant than the model implementation aspects. Reviewing research on non-English datasets for languages linguistically similar to Latvian reveals multiple projects with earlier methods, but the advancements of last two years described above do achieve improved results, and published work does yet not evaluate the effectiveness of these state of art approaches for non-English languages, so this needs experimental validation.

3 Task and evaluation dataset

The system was trained on and evaluated on the set of sentences used as corpus examples in the Latvian WordNet dataset, which have been manually linked to specific word senses and subsenses.

The sense inventory comes from the same dataset. It has a two-level granularity, listing senses which then may be split into subsenses. It's worth noting that the number of senses is substantially different for different words, with many rare words or terms having just one sense and no need for disambiguation, and some words having 10-20 subsenses grouped into five or more conceptually different senses.

For evaluation we selected 18 words covering the main parts of speech (7 verbs, 5 nouns, 3 adjectives and 3 adverbs) chosen out of the most frequent words in the corpus those that had multiple senses, were linguistically interesting, and had sufficient amount of annotated examples. 60% of the available annotated sentences were used to train the ”word expert” models and 40% of the annotated examples were used as test data for evaluation.

4 Model architecture

For initial proof of concept validation and testing of the data suitability a transformer-based deep learning model for word sense disambiguation was developed, very similar to (Hadiwinoto et al., 2019), pretraining on a large relevant corpus and fine-tuning for the classification of specific words.

The pretrained model used was a small (6 layers, 8 attention heads, 256 hidden unit size) version of BERT architecture trained on a combination from the Balanced Corpus of Modern Latvian (Levane-Petrova, 2019), Latvian Wikipedia and a web blog corpus, which is a reasonably diverse selection of approximately 50 million tokens. The small size of the pretrained model facilitates rapid experimentation as the model can be finetuned in a minute without the use of GPU.

A standard sequence classification architecture is used on top of this pretrained model, pooling the output and adding a single linear layer for the actual classification, with each word having a separate classification layer (”word expert”) with the number of classes determined by the number of subsenses in the dataset. The pretrained model is not updated during the training. This approach was chosen as the most popular approach seen in literature for integrating example sentence training data, which has shown good results for other languages. The technical implementation was done in PyTorch using the Huggingface transformers libraries.

5 Results and conclusions

Table 1 shows the results for classifying the annotated sense examples for a selection of 18 words. It's worth noting that these numbers are pessimistic as this selection focuses on highly ambiguous words with a large set of overlapping subsenses, excludes the less frequent “easy” words and the selection of examples overrepresents rare subsenses, so these numbers are not directly comparable to e.g. English WSD datasets. Accuracy is

Word	Accuracy main senses	Accuracy subsenses	Baseline main senses	Senses	Subsenses	Train samples	Test samples
domāt *to think* (verb)	49.1	49.1	31.9	6	9	174	116
dot *to give* (verb)	39.7	32.4	19.1	8	19	102	68
maksāt *to pay* (verb)	62.2	37.8	27.0	2	6	55	37
sēdēt *to sit* (verb)	45.6	24.6	15.8	4	10	84	57
sekot *to follow* (verb)	36.0	26.0	12.0	4	8	74	50
skriet *to run* (verb)	52.5	50.0	50.0	7	13	60	40
spēlēt *to play* (verb)	47.2	43.1	31.9	3	9	106	72
jautājums *question* (noun)	50.0	25.0	40.0	2	3	29	20
problēma *problem* (noun)	88.2	88.2	29.4	2	3	25	17
projekts *project* (noun)	80.6	74.2	32.3	3	4	45	31
sistēma *system* (noun)	47.5	45.0	20.0	4	6	60	40
zvaigzne *star* (noun)	65.8	63.2	28.9	3	4	55	38
liels *big* (adjective)	18.3	12.2	13.4	8	26	122	82
galvenais *main* (adjective)	61.1	41.7	44.4	4	7	52	36
iespējams *possible* (adjective)	50.9	33.3	24.6	3	6	84	57
daudz *much* (adverb)	55.9	52.9	41.2	3	4	51	34
vēl *more* (adverb)	39.7	22.4	19.0	4	6	87	58
vienmēr *always* (adverb)	70.6	70.6	70.6	2	2	24	17

Table 1. Accuracy of the proposed WSD model on a selection of 18 words

measured separately for the fine-grained subsense annotation and for the main senses of the word. The results achieve a significant improvement over a naive baseline of choosing the most popular sense in the training data.

The main result of this research is the development of the first proof of concept system of word sense disambiguation for Latvian, applying the latest state of art approaches which are very recent and have not yet been widely applied for languages similar to Latvian.

An immediate observation is the high variability of the accuracy for different words. It is plausible that this may reflect the relative distance between the senses, as for some words the senses may be fundamentally different and involve separate domains, while for others the difference may be a relatively narrow semantic change and because of this hard to disambiguate both for humans and automated systems.

The effect of number of samples and number of subsenses on accuracy is not obvious given the observed data. Each subsense in the dataset was allocated corpus examples for the primary needs of the dictionary, so words with more senses and subsenses also have more training and test examples.

A review of errors seems to indicate that in many cases, especially for the subsense distinction, a human would need a larger context than a single sentence in order to be certain about the proper interpretation. The current system works on a sentence basis, but in principle a longer context could be supplied.

The data seems to indicate that a larger set of senses is harder to disambiguate but it is not conclusive and it is plausible that the most significant factor is how different the specific senses are from each other. This should also align with how difficult it is for human annotators to assign senses, but this would need further research work.

6 Future work

An obvious future extension is to replace the currently used small transformer model with a model that is larger and has been pretrained on larger corpora. LVBERT (Znotins and Barzdins, 2020) is a possible candidate, but it is trained on a large news corpus which raises concerns about the omission of large classes of word senses such as colloquial language. It may be that a new transformed model would need to be trained on a more diverse corpus such as the recently updated Latvian National Corpora Collection (Saulīte et al., 2022) and experiment with the various updated transformer approaches that improve on the original BERT structure. We also received notice about a combined Lithuanian-Latvian pretrained model (Ulčar et al., 2021) that has the potential for improved results.

Another direction of future work is to prepare a representative evaluation corpus by annotating all word senses for a balanced set of running text. This data is required for proper evaluation to have a realistic frequency distribution, as rare examples are overrepresented in dictionary data.

Literature on English state-of-the-art suggests that substantial improvements can be achieved by integrating knowledge from the WordNet graph (Kumar et al., 2019; Bevilacqua and Navigli, 2020). This relies on a wide coverage sense graph, which is currently not available for Latvian, but there is potential that such a high-coverage graph could be developed soon through automated transfer of semantic links from the Princeton WordNet (Strankale and Stāde, 2022).

There seems to be potential to improve accuracy by directly applying more lexical data. There is work on using the “supersenses” from the WordNet hypernym ontology (Levine et al., 2019) and integrating synset gloss embeddings (Huang et al., 2019) as an additional data source for classification, so integrating gloss data would be a reasonable next step for extending the model. It's worth noting that these papers fully replace the supervised sentence examples, but combining the approaches could also yield useful results.

One of future applications for the developed system would be automatic disambiguation of Latvian corpora to enable corpus search for specific word senses, not only lemmas.

Acknowledgements

This research work is supported by the Latvian Council of Science, project "Latvian WordNet and Word Sense Disambiguation" (project No. LZP-2019/1-0464).

References

Barba, E., Pasini, T., Navigli, R. (2021). ESC: Redesigning WSD with extractive sense comprehension, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online, pp. 4661–4672.
`https://aclanthology.org/2021.naacl-main.371`

Bārzdiņš, G., Grūzītis, N., Nešpore, G., Saulīte, B., Auziņa, I., Levāne-Petrova, K. (2007). Ontological word sense disambiguation for discourse representation, *Proceedings of the 3rd Baltic Conference on Human Language Technologies*, pp. 33–40.

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP, *Linguistic Issues in Language Technology* **6**.

Bevilacqua, M., Navigli, R. (2020). Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2854–2864.

Blevins, T., Zettlemoyer, L. (2020). Moving down the long tail of word sense disambiguation with gloss informed bi-encoders, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 1006–1017.
`https://aclanthology.org/2020.acl-main.95`

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.
`https://aclanthology.org/N19-1423`

Hadiwinoto, C., Ng, H. T., Gan, W. C. (2019). Improved word sense disambiguation using pre-trained contextualized word representations, *arXiv preprint arXiv:1910.00194* .

Huang, L., Sun, C., Qiu, X., Huang, X. (2019). GlossBERT: BERT for word sense disambiguation with gloss knowledge, *arXiv preprint arXiv:1908.07245* .

Iacobacci, I., Pilehvar, M. T., Navigli, R. (2016). Embeddings for word sense disambiguation: An evaluation study, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 897–907.

Kumar, S., Jat, S., Saxena, K., Talukdar, P. (2019). Zero-shot word sense disambiguation using sense definition embeddings, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5670–5681.

Levane-Petrova, K. (2019). Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos, *Language: Meaning and Form* **10**, 131–146. The Balanced Corpus of Modern Latvian, its role in grammar studies.
`https://www.apgads.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/Valoda-nozime-forma/VNF-10/vnf_10-12_Levane_Petrova.pdf`

Levine, Y., Lenz, B., Dagan, O., Ram, O., Padnos, D., Sharir, O., Shalev-Shwartz, S., Shashua, A., Shoham, Y. (2019). Sensebert: Driving some sense into BERT, *arXiv preprint arXiv:1908.05646* .

Lokmane, I., Rituma, L., Stāde, M., Klints, A. (2021). The Latvian WordNet and word sense disambiguation: Challenges and findings, *Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography* p. 76.

Paikens, P., Grasmanis, M., Klints, A., Lokmane, I., Pretkalniņa, L., Rituma, L., Stāde, M., Strankale, L. (2022). Towards Latvian WordNet, *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC).*

Pasini, T., Raganato, A., Navigli, R. et al. (2021). Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation, *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI Press.

Saulīte, B., Darģis, R., Grūzītis, N., Auziņa, I., Levāne-Petrova, K., Pretkalniņa, L., Rituma, L., Paikens, P., Znotiņš, A., Strankale, L., Pokratniece, K., Poikāns, I., Bārzdiņš, G., Skadiņa, I., Baklāne, A., Saulespurēns, V., Jānis, Z. (2022). Latvian national corpora collection – korpuss.lv, *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC).*

Song, Y., Ong, X. C., Ng, H. T., Lin, Q. (2021). Improved word sense disambiguation with enhanced sense representations, *Findings of the Association for Computational Linguistics: EMNLP 2021*, Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 4311–4320.
`https://aclanthology.org/2021.findings-emnlp.365`

Strankale, L., Stāde, M. (2022). Automatic word sense mapping from princeton WordNet to latvian WordNet, *Proceedings of the 14th International Conference on Agents and Artificial Intelligence - Volume 1: NLPinAI,*, INSTICC, SciTePress, pp. 478–485.

Ulčar, M., Žagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages, *arXiv preprint arXiv:2107.10614* .

Vial, L., Lecouteux, B., Schwab, D. (2019). Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation, *arXiv preprint arXiv:1905.05677* .

Znotins, A., Barzdins, G. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding, *Human Language Technologies - The Baltic Perspective*, Vol. 328, IOS Press, pp. 111–115.
`http://ebooks.iospress.nl/volumearticle/55531`

Received August 19, 2022 , accepted August 24, 2022