A Unified Sense Inventory for Word Sense Disambiguation in Polish

Arkadiusz Janz[0000−0002−9203−5520], Agnieszka Dziob[0000−0002−7425−8181], Marcin Oleksy[0000−0001−7740−5557], and Joanna Baran[0000−0001−6792−7028]

Wrocław University of Science and Technology
Department of Computational Intelligence, Wrocław, Poland,
{arkadiusz.janz|joanna.baran}@pwr.edu.pl

Abstract. We introduce a comprehensive evaluation benchmark for Polish Word Sense Disambiguation task. The benchmark consists of 7 distinct datasets with sense annotations based on *plWordNet*–4.2. As far as we know, our work is a first attempt to standardise existing sense annotated data for Polish. We also follow the recent trends of neural WSD solutions and we test transfer learning models, as well as hybrid architectures combining lexico-semantic networks with neural text encoders. Finally, we investigate the impact of bilingual training on WSD performance. The bilingual model obtains new State of the Art performance in Polish WSD task.

Keywords: WSD · Knowledge bases · Neural models · Benchmarking

1 Introduction

For over 50 years, the interest in Word Sense Disambiguation (WSD) has not changed. A great progress has been made on improving general quality of WSD data and algorithms. Still, WSD is an open problem for low-resource languages due to the lack of sense annotated datasets. Modern WSD algorithms require great volumes of data to successfully disambiguate word meanings in a multi-domain setting. On the contrary, multilingual language models and transfer learning methods have been proved to be quite effective when annotated data is unavailable [15].

Polish WSD data seemed to be quite scarce, especially for lemmas in verb and adjective categories. This issue has always been seen as a limiting factor for Polish language, slowing the development of supervised WSD approaches. On the other hand, *plWordNet* with its lexico-semantic structure and sense descriptions has been rebuilt and significantly extended with more data [4, 19, 8, 22]. One of the main issues is that, until now, Polish WSD data has resided in various sources and formats. In this paper, we decided to explore all necessary requirements to develop an effective WSD solution for Polish by following a data-centric approach. The main contributions of this paper are as follows:

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_70

- We cleaned, updated, and substantially extended existing sense annotated corpora, and we unified them under one modern knowledge-base – *plWordNet*–4.2. We introduced 2 manually annotated WSD datasets of significant size. Overall, 7 distinct datasets covering more than 60% of senses from Polish wordnet have been collected. Finally, we propose a first WSD evaluation framework for Polish in one package with all resources integrated with Princeton WordNet [12] and BabelNet [14] indices.
- We evaluated crosslingual architectures as well as monolingual neural models and hybrid solutions on Polish data at scale. Such evaluation has been so far impossible due to the lack of the necessary linguistic resources and annotated data of considerable size. We selected two modern neural architectures to train for WSD task: XL-WSD solution [15] and EWISER [2].
- We investigated bilingual training and its impact on Polish WSD for the first time. We show that bilingual training is not necessarily a harmful process [21]. The investigated models raised Polish WSD performance to a new level.

2 Related Work

Resources. Princeton WordNet (PWN) [12] is a lexical database used in various NLP applications. It was also a foundation for wordnets created in other languages, often as a translation from English. The eXtended WordNet project [6] aimed to enrich PWN with information found in sense definitions and usage examples, on the basis of which Princeton WordNet Gloss Corpus [1] was later developed. Words extracted from the definitions (also called "glosses") in synsets were manually linked to their context-appropriate sense in WordNet. This created a major training dataset for the WSD task for a long time. Another extension of PWN was proposed in [17] where the authors utilised external knowledge resources such as *Wikipedia* to increase the general density of semantic relations in wordnet, leading to higher WSD performance. With this assumption the BabelNet [14] was created, taking the role of main WSD resource.

Raganato et al. [18] introduced a unified WSD evaluation framework for English. It merged all Senseval and SemEval data into a single dataset containing 7,253 instances to disambiguate. Pasini et al. [15] proposed XL-WSD – a crosslingual dataset for the WSD task. It stands as a key semantic benchmark not only in terms of size, but also in terms of coverage. Unfortunately, it does not contain any test data for Polish. Leveraging the resources of Open Multilingual WordNet [20] and BabelNet [14] allowed to create new multilingual evaluation benchmarks.

The work on the creation of Polish WSD corpora began about 10 years ago. `KPWr` [3] and `Składnica` [10] were the first evaluation resources for Polish WSD. However, they were developed at the time of the official announcement of *plWordNet*-2.1. In [7] the development of sense annotated data included partial

[1] `https://wordnetcode.princeton.edu/glosstag.shtml`

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_70

update of Składnica and KPWr data to *plWordNet*-3.2. Still, the available WSD resources were of moderate size.

Methods. We can distinguish three main approaches to WSD classification task: *knowledge-based*, *supervised* and *hybrid* combining the previous two. Despite their large coverage of senses, the knowledge-based approaches are easily outperformed by supervised solutions when sufficient training data is provided [18]. The key factor of their performance is based on assumption that sense descriptions in the knowledge base are reflecting the natural context of sense occurrences in the corpora. This applies to both textual descriptions as well as their lexico-semantic structure in the knowledge-graph. [10, 9] represent the adaptations of knowledge-based methods such as [13, 1] to Polish language. Recently, with the growing popularity of deep neural language models, new architectures such as EWISER [2] (*hybrid*) or XL-WSD [15] (*supervised*) have been proposed. They proved that the multilingual models can be successfully used to prepare an effective solution for languages other than English.

3 Polish WordNet as a knowledge base

plWordNet (pol. *"Słowosieć"*; *plWN*) is a large wordnet of Polish built from scratch and manually mapped onto PWN [19], in which the central semantic entity is the *lexical unit* (LU). It is integrated with SUMO ontology [16], Wikipedia [11], Polish valency lexicon Walenty [5], and enriched with an extensive emotive annotation [22]. Starting from version *4.2* (see Table 1) (2020) all efforts have focused on increasing the density of lexico-semantic structure and revising sense granularity. With all of its extensions, plWordNet could possibly improve WSD performance in other languages when integrated properly with WSD models.

Table 1: Descriptive statistics of wordnet-based sense inventories for Polish. The average polysemy ratio was computed for polysemous lemmas only.

Feature	***plWN 2.1***	***plWN 3.2***	***plWN 4.2***
`LU`	206 567	286 804	294,842
Multi-word `LU`	53 752	70 019	71 133
Synsets	151 252	221 101	227 369
`LU` with gloss or usage example	37 207	145 901	155 290
Average length of utterance	12.56	11.54	11.08
Average number of senses per lemma	2.79	2.96	3.05

4 Polish WSD Inventory

Each corpus was prepared according to the following data pre-processing pipeline. All texts were segmented, tokenized and underwent morpho-syntactic analysis. We added an additional annotation layer with automatically recognized multiword expressions (MWE) existing in plWordNet [2]. Most of our datasets were

[2] `https://clarin-pl.eu/dspace/handle/11321/508`

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_70

annotated in 2+1 system - two linguists working independently supported by third super-annotator to resolve inconsistencies. Lastly, the tokens representing open-class words were selected for manual sense correction or annotation. All corpora described below were manually updated to plWN 4.2 [3].

`Składnica (SK)` is a sense-annotated treebank [5] used in the past as an evaluation set for knowledge-based WSD approaches for Polish [10]. Re-introduced at *PolEval's WSD competition Task 3* [7], recently has been trending as a training set. The sentences in `Składnica` were carefully parsed and manually annotated. `KPWr-N` annotation was based on a lexical sampling approach for a small set of words [3, 10] – here we present an updated version with manually extended semantic annotation. `Sherlock Holmes: The Adventure of The Speckled Band (SPEC)` by Sir Arthur Conan Doyle has been translated to Polish by a team of professionals as a part of The NTU Multilingual Corpus [20], and manually tagged both with morphological information and WSD. Unlike the first annotation of the `KPWr-N` corpus, in `KPWr-100` the process was aimed at full-text sense annotation. Documents from various sources and representing different functional styles and genres were manually tagged. `SPEC` and `KPWR-100` were introduced in [7] as a test framework for the competition – in this work we updated all of its sense annotations from *plWordNet*–3.2 version to 4.2.

Table 2: The overall number of tokens and lemmas in our corpora. We also provide a percentage of covered senses with respect to *plWordNet*–4.2.

Feature	**Dataset**						
	SK	SPEC	EmoGLEX	WikiGLEX	KPWr-N	KPWr-100	GLEX
All tokens	136 075	9 087	290 313	96 308	438 505	32 522	5 020 817
All lemmas	17 065	2 212	22 927	13 996	31 142	6 317	170 200
Sense coverage	5.58%	0.85%	4.53%	3.49%	1.31%	2.30%	60.53%
Avg. polysemy in corpus	2.6	2.3	2.6	2.5	2.9	2.5	–
Avg. polysemy in plWN	4.2	5.1	4.2	4.3	3.9	4.5	–
Sense annotations	43 776	3 947	N/A	N/A	14 429	14 004	333 254
After mapping	31 294	3 113	N/A	N/A	10 603	11 334	292 119

`GLEX` is a corpus of glosses and usage examples. It includes synsets which contain the lexical units having at least one natural language uterrance. Among the whole corpus, we distinguished two distinct subcorpora with full-text sense annotation. The first one, `WikiGLEX`, contains glosses and usage examples of senses that represent the intersection of `plWordNet` and Wikipedia. The second one, `EmoGLEX`, was created from synsets with at least one lexical unit containing sentiment annotation [22] and emotive examples obtained in [8] project. The aforementioned corpora were cleaned, accurately annotated and compiled together within our framework.

[3] `https://clarin-pl.eu/dspace/handle/11321/891`

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_70

5 Evaluation

We use the datasets described in section 4 as a basis for our evaluation framework. We decided to designate default data splits for training and evaluation. Sense distribution and their coverage are usually seen as key WSD factors to consider when building representative training set. We investigated the following scenarios.

1. **Zero-shot** setting evaluated multilingual language models fine-tuned on English WSD data only. Training set consisted of `SemCor` mixed with `PWN` definitions and usage examples. We adapted XLM RoBERTa Large to the task as it was proposed in XL-WSD.
2. **Monolingual** approach was focused on assessing monolingual models as they are expected to perform better than transfer learning when sufficient amount of data is available. To prepare this model we trained EWISER architecture on `GLEX` corpus only due to its large vocabulary and sense coverage. The remaining available Polish corpora were used in evaluation step.
3. **Bilingual** setting examined the impact of bilingual training on Polish WSD performance. We trained the same model as in *monoligual* scenario on extended corpora consisting of SemCor data, PWN Gloss Corpus and our `GLEX` corpus.

Parameter settings. The models and their parameters were tuned on validation data using early stopping strategy with validation loss as a core metric. The number of epochs was set to 30. In the zero-shot setting, SemEval's 2015 dataset was used as a validation set following XL-WSD research. In other settings, we used a sample of GLEX corpora as our validation data.

Data preprocessing. To train and evaluate the models we mapped all of sense annotations in our corpora onto BabelNet as the EWISER architecture requires all of the data to be compatible with BabelNet indices. This mapping was done by taking existing interlingual links between plWordNet and PWN including *i-synonyms*, *i-hypernyms*, *i-hyponyms*, preferring *i-synonyms* at first place. We also prepared a joint sense inventory with Polish and English lemmas and their candidate meanings mapped onto BabelNet indices.

As a baseline solution, we considered WoSeDon – a knowledge-based model with PageRank algorithm at its core [10, 1]. Table 3 presents a summary of our experimental part. We can notice that the baseline architecture was outperformed by all of tested neural architectures. The zero-shot model has proved to be quite effective on frequent-sense data such as `SPEC`, `SK` and `KPWr-100`. However, for non-trivial data e.g. `KPWr-N` with diverse sense distribution the results are less optimistic. As expected, the monolingual model performs slightly better than zero-shot solution. The bilingual solution has achieved the greatest performance among all models being evaluated. However, the results for `KPWr-N` dataset suggest that a monolingual model might be better when dealing with rare senses.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_70

Table 3: F1-scores of tested architectures. Asterisk represents development data.

Architecture	Datasets				
	SK	SPEC	KPWr-N	KPWr-100	GLEX
WoSeDon	—	62.30	—	64.72	*
Zero-shot (XLMR-L)	70.07	73.07	47.81	70.24	68.31
Monolingual (EWISER)	70.42	72.05	51.52	70.51	*
Bilingual (EWISER)	72.52	75.55	50.41	72.48	*

6 Conclusions

We proposed a new evaluation benchmark for Polish WSD task. We summarized all of its resources including newly obtained corpora as well as the knowledge base used as its sense inventory. We evaluated modern language models on our benchmark data achieving a new State-of-the-Art performance in Polish WSD. The results suggest that neural models can be successfully utilised to prepare a good enough WSD solution for Polish. Still, the evaluation for rare senses uncovers the main issue of existing data sources – the most frequent sense bias. In further work we plan to improve language models in terms of their adaptability to new and rare senses. We release presented data and code for public use (available at `https://github.com/CLARIN-PL/polish-wsd-datasets`).

Acknowledgments

This work was co-financed by (1) the Polish Ministry of Education and Science, CLARIN-PL; (2) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, project number POIR.04.02.00-00C002/19; and (3) by the National Science Centre, Poland, grant number 2018/29/B/HS2/02919.

References

[1] Agirre, E., López de Lacalle, O., Soroa, A.: The risk of sub-optimal use of open source NLP software: UKB is inadvertently state-of-the-art in knowledge-based WSD. In: Proc. of Workshop for NLP Open Source Software (NLP-OSS). Melbourne, Australia (2018)

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

[3] Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., Wardyński, A.: KPWr: Towards a free corpus of Polish. In: Proc. of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey (May 2012)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_70

[4] Dziob, A., Piasecki, M., Rudnicka, E.K.: plwordnet 4.1 – a linguistically motivated, corpus-based bilingual resource. In: Proc. of the 10th Global Wordnet Conference. pp. 353–362
[5] Hajnicz, E.: Lexico-semantic annotation of składnica treebank by means of PLWN lexical units. In: Proc. of the 7th Global Wordnet Conference. Tartu, Estonia (Jan 2014)
[6] Harabagiu, S., Moldovan, D.: Knowledge processing on an extended wordnet. WordNet: An electronic lexical database **305**, 381–405 (1998)
[7] Janz, A., Chlebus, J., Dziob, A., Piasecki, M.: Results of the poleval 2020 shared task 3: Word sense disambiguation. Proc. of the PolEval 2020 Workshop p. 65
[8] Janz, A., Kocon, J., Piasecki, M., Zasko-Zielinska, M.: plwordnet as a basis for large emotive lexicons of polish. Proc. of Human Language Technologies as a Challenge for Computer Science and Linguistics Poznan pp. 189–193 (2017)
[9] Janz, A., Piasecki, M.: Word sense disambiguation based on constrained random walks in linked semantic networks. In: Proc. of the International Conference on Recent Advances in Natural Language Processing. Varna, Bulgaria (2019)
[10] Kędzia, P., Piasecki, M., Orlińska, M.: Word sense disambiguation based on large scale polish clarin heterogeneous lexical resources. Cognitive Studies (15) (2015)
[11] Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S., Kędzia, P.: plwordnet 3.0–a comprehensive lexical-semantic resource. In: Proc. of COLING 2016. pp. 2259–2268 (2016)
[12] Miller, G.A.: WordNet: An electronic lexical database. MIT press (1998)
[13] Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics **2** (2014)
[14] Navigli, R., Ponzetto, S.P.: Babelnet: Building a very large multilingual semantic network. In: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 216–225 (2010)
[15] Pasini, T., Raganato, A., Navigli, R.: Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In: Proc. of the AAAI Conference on Artificial Intelligence. AAAI Press (2021)
[16] Pease, A.: Ontology - A Practical Guide. Articulate Software Press (2011)
[17] Ponzetto, S.P., Navigli, R.: Knowledge-rich word sense disambiguation rivaling supervised systems. In: Proceedings of the 48th annual meeting of the association for computational linguistics. pp. 1522–1531 (2010)
[18] Raganato, A., Camacho-Collados, J., Navigli, R.: Word sense disambiguation: A unified evaluation framework and empirical comparison. In: Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 1. pp. 99–110 (2017)
[19] Rudnicka, E., Maziarz, M., Piasecki, M., Szpakowicz, S.: A strategy of mapping polish wordnet onto princeton wordnet. In: Proc. of COLING 2012. pp. 1039–1048 (2012)
[20] Tan, L., Bond, F.: Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). In: Proc. of the 25th Pacific Asia Conference on Language, Information and Computation. Singapore (2011)
[21] Wang, Z., Lipton, Z.C., Tsvetkov, Y.: On negative interference in multilingual models: Findings and a meta-learning treatment. In: Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). p. 4438—4450 (2020)
[22] Zaśko-Zielińska, M., Piasecki, M.: Towards emotive annotation in plwordnet 4.0. In: Proc. of the 9th Global Wordnet Conference. pp. 153–162 (2018)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_70