WordNet-based Data Augmentation for Hybrid WSD Models

Arkadiusz Janz and Marek Maziarz
Wrocław University of Science and Technology
{arkadiusz.janz|marek.maziarz}@pwr.edu.pl

Abstract

Recent advances in Word Sense Disambiguation suggest neural language models can be successfully improved by incorporating knowledge base structure. Such class of models are called hybrid solutions. We propose a method of improving hybrid WSD models by harnessing data augmentation techniques and bilingual training. The data augmentation consist of structure augmentation using interlingual connections between wordnets and text data augmentation based on multilingual glosses and usage examples. We utilise language-agnostic neural model trained both with SemCor and Princeton WordNet gloss and example corpora, as well as with Polish WordNet glosses and usage examples. This augmentation technique proves to make well-known hybrid WSD architecture to be competitive, when compared to current State-of-the-Art models, even more complex.

1 Introduction

Word Sense Disambiguation is well recognised issue in Natural Language Processing. Due to word ambiguity it is impossible to give *a priori* a proper semantic interpretation of a text, so senses ought to be disamiguated. In recent years a great improvement has been achieved in the field with the use of deep neural networks (DNN). For low-resourced languages, however, WSD is still an open problem because of the lack of large-scale sense annotated corpora required by modern neural models.

Large number of categories (which are senses themselves) makes the task very hard for DNN classifiers, because of the bottleneck of sense annotation sparseness. Constructing a large sense annotated corpus is a very laborious task, so this problem affects NLP for most world languages (the estimated number of which exceeds 6,000). On the other hand, even NLP for languages that possess vast WSD corpora (i.e. SemCors and extensive wordnet-based corpora) has to cope with a huge number of senses that are rarely occurring in texts (for such senses the available DNN representation might not be sufficient).

Two main solutions have been proposed to these problems: first, the usage of knowledge bases facilitates WSD algorithm through propagating information within a semantic network. Second, the use of pre-trained language models, especially multilingual (or language agnostic) allows to train a model on existing resources (especially English ones) and apply it to a new language context.

We present a slight but successful modification of the EWISER model (Bevilacqua and Navigli, 2020a) in which we merge both approaches. The novelty lies in special data augmentation technique focused on structural properties of knowledge bases in other than English language, namely Polish. Starting from EWISER language-agnostic architecture pre-trained on English and Polish sense annotated datasets, we then propagate DNN vector representations through combined structures of Princeton WordNet and Polish Wordnet, two largest nowadays wordnets in the world. This modification boost the WSD multilingual performance above current State-of-the-Art solutions based on multilingual language models e.g. XL-WSD framework (Pasini et al., 2021), and gives comparable behaviour to earlier SOTA model of CONSEC, despite the fact that EWISER architecture - even with our modifications - is much simpler.

2 Related Work

The supervised approaches have proved to be the most effective solution to WSD when a representative training sample is available. With recent progress in neural language modeling the supervised solutions have been improved even more and outperformed earlier models on almost every single benchmark. However, the existing WSD data yet has its flaws, including a non-representative training sample for verb, adverb and adjective senses,

most frequent sense bias, and limited sense coverage. Although very successful, the supervised models are overfitting easily to training samples which harms their generalisation abilities and reduces sense coverage when non-representative samples are used for training (Kumar et al., 2019; Bevilacqua and Navigli, 2020a). The knowledge-based solutions were designed to increase the coverage of underrepresented word senses when a limited training sample is available. However, the performance gap between supervised and knowledge-based solutions encouraged the researchers to focus more on former approaches. The prior work on supervised models considered WSD task as token classification problem where the model learns to generate discrete labels representing predicted meanings (Iacobacci et al., 2016; Raganato et al., 2017; Popov, 2018). A typical architecture consisted of neural context encoder and sense discrimination layer e.g. LSTM with attention and softmax layer trained on SemCor data to disambiguate tokens in a fully supervised manner.

Recent studies in the area of Word Sense Disambiguation show that the most successful solutions are based on hybrid architectures with a strong emphasis on zero-shot supervision. A zero-shot component was introduced to replace full supervision and improve the ability of generalising to unseen senses (Kumar et al., 2019). Subsequent approaches utilised the benefits of transformer architectures (Huang et al., 2019; Du et al., 2019) and representation learning using external knowledge sources, such as sense definitions (Luo et al., 2018; Huang et al., 2019; Blevins and Zettlemoyer, 2020) and sense usage examples. On the other hand, structural properties of lexico-semantic networks used to be ignored in neural architectures. Recent studies show that hybrid solutions utilising textual descriptions of senses together with their structural properties can also improve WSD performance.

Most related to our work is XL-WSD framework with a crosslingual benchmark built on the basis of Open Multilingual WordNet data and BabelNet resources. The benchmark has been introduced as a platform to evaluate zero-shot WSD methods and crosslingual transfer with multilingual language models. Other multilingual solutions include MULAN (Barba et al., 2021a), EWISER (Bevilacqua and Navigli, 2020a), CONSEC (Barba et al., 2021b). However, only few of them were actually evaluated against all of datasets available in XL-WSD framework. The usual crosslingual evaluation setting consists of English, Spanish, French, German and Italian datasets proposed at SemEval competition. XL-WSD was a step towards preparing a crosslingual evaluation at scale including more languages. As far as we know, none of the previous solutions evaluated within XL-WSD framework were hybrid models joining neural text encoders with structural knowledge base features.

Regarding the Negative Transfer phenomenon, several studies were focused on identification of troublesome NLP tasks where simultaneous fine-tuning of multilingual language models to downstream tasks has a harmful impact on model performance (Wang et al., 2020). However, none of them were focused strictly on WSD task. It is an open issue whether Negative Transfer occurs when fine tuning multilingual language models to WSD task.

3 Resources

3.1 XL-WSD Framework

Pasini et al. (Pasini et al., 2021) prepared a framework of gold-standard resources for testing WSD models for 17 languages and English. They started from a sense inventories created on the basis of a version of Open Multilingual Wordnet (OMW) (Bond and Paik, 2012), and the extended version of OMW (based on Wiktionary data sets) (Bond and Foster, 2013). OMW identifiers are simply PWN synset IDs, so a new sense is announced each time a lemma is ascribed a new PWN synset. The sense inventories are obtainable online.[1] Princeton WordNet synset IDs were translated to BabelNet internal identifiers for authors' convenience. The authors pre-trained multilingual language model based on XLM-RoBERTa architecture (Conneau et al., 2020) to assess cross-lingual transfer capabilities of these models in a word sense disambiguation task. We made use of XL-WSD inventories of 14 languages (excluding Italian, Japanese and Korean due to sense inventory issues and missing senses discovered in XL-WSD framework).

Our models were trained on Princeton WordNet glosses and usage examples, as well as on SemCor and tested on SemEval tasks and texts (glosses and usage examples) from several wordnets. Table 1 describes the data sets in terms of annotated text origin (as either wordnet-based or SemEval-based).

[1] https://sapienzanlp.github.io/xl-wsd/

Language	Type	#Instances
en	SemEval	8 062
bg	WN-based	9 968
ca	WN-based	1 947
da	WN-based	4 400
de	SemEval	862
es	SemEval	1 851
et	WN-based	1 999
eu	WN-based	1 580
fr	SemEval	1 160
gl	WN-based	2 561
hr	WN-based	6 333
hu	WN-based	4 428
nl	WN-based	4 400
sl	WN-based	2 032
zh	WN-based	9 568

Table 1: Language-specific test sets, their type and size as reported in (Pasini et al., 2021) publication. SemEval datasets usually are easier to disambiguate when compared against WN-based datasets.

Link type	Count
i-hyponyms	181 029
i-hypernyms	181 032
i-synonyms	93 654
Total	455 715

Table 2: Number of interlingual connections between plWordNet-3.2 and Princeton WordNet by category.

3.2 Polish Data

Polish WordNet (plWN) was heavily inter-linked with Princeton WordNet (Rudnicka et al., 2012). More than two hundred thousand relation instances were used linking Polish-English counterpart synsets, among which inter-lingual synonymy, inter-lingual hyponymy and inter-lingual hypernymy were the most prominent. In Table 2 we present newest statistics concerning the manual mapping (Dziob et al., 2019). We used the mapping in the process of augmenting the structure of PWN with new links (see Sec. 4.1 below for details).

4 Models

As a baseline architecture we decided to use EWISER (Bevilacqua and Navigli, 2020b) as its codebase is extensible and freely available. EWISER is a supervised hybrid architecture utilising sense annotated corpora and knowledge base structure simultaneously. The model is based on transformer architecture with additional sense discrimination layer and structured logit mechanism injecting structural information into model during training. The key idea is to utilise existing wordnet links between senses to reinforce training procedure and incorporate logit scores of neighboring senses into scoring function of word's candidate meanings.

4.1 Augmenting the Structure

We augmented Princeton WordNet, PWN (Fellbaum, 1998), structure with semantic relations obtained from Polish WordNet, plWN (Maziarz et al., 2016) in the following manner:

Consider two pairs of counterpart synsets from plWN and PWN $s_1^{plWN} \overset{I-rel}{\longleftrightarrow} s_1^{PWN}$ and $s_2^{plWN} \overset{I-rel}{\longleftrightarrow} s_2^{PWN}$, where "*I*-rel" signifies an inter-lingual relationship. Each time when there exists a short path between the two Polish synsets in plWN, we add a new link: $s_1^{PWN} \leftrightarrow s_2^{PWN}$ to PWN. We assumed that for synonymous counterparts the distance should not exceed 2, while for homonymous counterparts the maximum path length was set to 1.

The above assumptions were fulfilled with simple matrix algebra. Let's talk about separate sets: (i) I^{hyp} of all plWN synsets that have their *I*-hypernyms or *I*-hyponyms on the PWN side and (ii) I^{syn} of all plWN synsets that have their *I*-synonyms in PWN.

(i) For the I-hyponymy/I-hypernymy case the procedure is straightforward. We simply took the original adjacency plWN matrix A and filter it leaving only synsets from the set I^{hyp}, i.e. $H = \{a_{ij}\}_{i,j \in I^{hyp}}$.

(ii) For the I-synonymy case we started from the plWN adjacency matrix A and took its square $S = A^2$ (i.e. the matrix product of 2 copies of A). Its elements $\{s_{ij}\}$ are indexed by synset identifiers i, j and represent the number of random walks of length 2 on the plWN graph (Kranda, 2011). Calculating $S' = \{\operatorname{sign} s_{ij}\}$, i.e. setting non-zero elements of the matrix to 1, and adding $A + (S' - \mathbb{I}) = M = \{m_{ij}\}$, we get a matrix with new adjacency links (representing the distance of 2 or less steps in the original graph A). Out of the matrix M we construct the new matrix E with picking up only those synsets that are in the set

I^{syn}, i.e. $E = \{m_{ij}\}_{i,j \in I^{syn}}$.

Taking into account all relationships obtainable from matrices H and E we finally land with the set of new links to be added to PWN.

4.2 Augmenting the Data

Nearly 146,000 Polish synsets are described by a gloss and/or by (a) usage example(s). These samples were used to extend EWISER's training data. To obtain their textual descriptions we used interlingual links from plWordNet 3.2 including interlingual synonymy, hyponymy and hypernymy.

In (Pasini et al., 2021) authors used machine translated PWN glosses and usage examples and found no significant improvement over other models. In contrast to their approach, we used Polish glosses and native natural language examples avoiding translation disadvantages (see Sec. 4.3 below for details).

4.3 Bilingual Training

To investigate the impact of bilingual training on WSD performance we built a mixed sense inventory consisting of Polish and English lemmas with their candidate meanings. To create this inventory we used interlingual mapping between Polish and English wordnet meanings, mainly synonymy, hypernymy and hyponymy links. We believe multilingual downstream task fine-tuning might be beneficial for tasks such as WSD, since it is strongly interconnected with training procedure of multilingual language models (usually on parallel corpora), e.g. multilingual MLM in XLM-R. However, for tasks such as POS tagging or NER recognition issues such as Negative Transfer (also called Negative Interference) model performance is decreased during multilingual training (Wang et al., 2020). Thus our work is one of the first attempts to investigate Negative Transfer phenomenon in WSD task.

5 Experiments

In this section we present the results of our experimental part. We decided to split evaluation into two different settings. First, we would like to investigate the impact of underlying language model on WSD performance. The second setting is focused on data augmentation using plWordNet data (the network structure, as well as glosses and examples).

5.1 Settings

The authors of EWISER in their original work integrated their architecture with mBERT language model (Devlin et al., 2019). However, recent progress on multilingual language modeling brought new and more effective language models such as XLM-RoBERTa (Conneau et al., 2020), T5 (Raffel et al., 2020), mBART (Liu et al., 2020). The XLM architecture is oftenly choosed as a main language model for various downstream tasks. It was also the basis for crosslingual evaluation of zero-shot solutions within XL-WSD framework. However, as far as we know, the XLM architecture has never been evaluated within hybrid WSD approaches. Thus, in our first setting we evaluate the EWISER architecture with XLM-RoBERTa-Large model as underlying context encoder.

In second setting we focused mainly on the proposed data augmentation methods – structure expansion and corpora expansion. We investigate the impact of Polish data on WSD performance in English as well as in multilingual setting with multiple languages. The first baseline solution utilises a zero-shot architecture proposed in XL-WSD framework with XLMR-Large model. Contrary to EWISER, this architecture is not a hybrid solution and does not utilise structural properties of knowledge bases. We split this experiment into two parts. The first part is focused on structure augmentation using interlingual synonymy and relation propagation over wordnet. The second part of this setting evaluates a joint model where the structure augmentation technique is combined with additional sense data including glosses and sense utterances. A bilingual dataset and bilingual sense inventory are used to train the joint model.

5.2 Hyperparameter Tuning

The hyperparameters were finetuned using a preselected validation set. We chose SemEval 2015 data set as our development data following the way it was used in the literature. We applied early stopping procedure to prevent the models from overfitting to training data, as it was proposed in (Bevilacqua and Navigli, 2020b). The experiments were repeated at least 5 times for each model.

6 Results and Discussion

In tests on 15 languages our technique turned out to be successful in beating the XL-WSD and the EWISER model and comparable to some extent

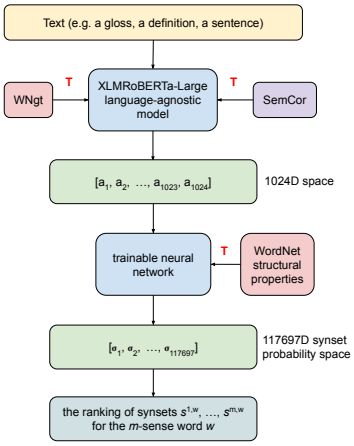

Figure 1: The DNN architecture of EWISER. We provided Polish language data both for XLM-RoBERTa language model (plWordNet glosses and usage examples) and for the output neural network layer (new relation instances for Princeton WordNet derived from plWordNet).

with the CONSEC model. Table 3 illustrates multilingual performance of all models, as compared with baselines - EWISER, CONSEC[2] and XLM-RoBERTa from XL-WSD framework.

Since testing data sets were constructed independently, we decided to compare average model F1 performances. *U*-Mann-Whitney paired test was applied to the task, separately for CONSEC and for XL-WSD with EWISER) and *p*-values were corrected for false discovery ratio through Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Our two models performed better *on average* than XL-WSD (XLMR-L) and EWISER baseline models (for 15 languages) and not worse than CONSEC model (for 6 languages).

Presented in this paper experiments proved that augmenting English training data sets with glosses and examples from other than English wordnet can lead to the improvement of a multilingual WSD algorithm. The proposed novel technique of augmenting Princeton WordNet structure also resulted in better than or equal to SOTA scores. Surprisingly, used here EWISER architecture is simpler than current SOTA DNN models. This suggests the validity of training data enlargement and curation techniques. The step that could not be fully superseded by constructing new, even more sophisticated DNN architectures.

In the future we plan to investigate new ways of enriching Princeton WordNet structure with relation instances derivable from Polish WordNet network. Since we utilised only separate sets of *I*-synonyms and *I*-hyponyms/*I*-hypernyms, it is obvious that these two types of bilingual counterparts could be treated jointly. For instance, we may link in PWN an English *I*-synonym with an English *I*-hyponym, if a path is not too long. This enrichment will provide us with new, high quality relations. Also testing different path lengths via plWordNet is planned.

[2]The evaluation of CONSEC model was limited to the results provided by the authors in (Barba et al., 2021b). At the time of publication, the training procedure was not fully reproducible and the codebase was incompatible with XL-WSD sense indices.

Acknowledgments

This research was financed by the National Science Centre, Poland, grant number 2018/29/B/HS2/02919, and supported by the Polish Ministry of Education and Science, Project CLARIN-PL.

References

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2021a. Mulan: Multilingual label propagation for word sense disambiguation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3837–3844.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021b. Consec: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Michele Bevilacqua and Roberto Navigli. 2020a. 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*, pages 2854–2864.

Michele Bevilacqua and Roberto Navigli. 2020b. 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*, pages 2854–2864.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

Language ISO 639-1	Baselines EWISER [e]	CONSEC [c]	XLM-R [x]	EWISER-augmented +PLWN (Es)	+PLWN (Es+Ts)
en$^+$	78,9	**83.4**	76.3	79.9	79.6
bg	74,2	—	72.0	74.7	**75.4**
ca	53,6	—	50.0	54.2	**55.2**
da	82,6	—	80.6	82.8	**83.3**
de$^+$	83,1	**84.2**	83.2	83.1	82.9
es$^+$	77,0	77.4	75.8	77.4	**78.2**
et$^+$	71,1	69.8	66.1	70.9	**71.5**
eu	50,2	—	47.2	50.5	**50.8**
fr$^+$	83,8	84.4	83.9	83.9	**84.7**
gl	**67,7**	—	66.3	66.4	67.4
hr	74,1	—	72.3	74.2	**74.3**
hu	**73,7**	—	67.6	73.6	**73.7**
nl$^+$	63,2	63.3	59.2	63.5	**64.1**
sl	66,6	—	68.4	**68.0**	67.5
zh	56,1	—	51.6	56.3	**56.5**
mean$^+$	76.1	**77.0**	74.1	76.5	76.8
mean	70.3	—	68.0	70.6	**71.0**
median$^+$	77.9	**80.4**	76.1	**78.5** $^{[c]}$ (=)	**78.6** $^{[c]}$ (=)
median	73.6	—	68.4	**73.6** $^{[e]}$ * (↑) $^{[x]}$ ** (↑)	**73.7** $^{[e]}$ ** (↑) $^{[x]}$ *** (↑)

Table 3: Multilingual performance of different models in terms of F1 scores. Symbols: "+PLWN" – Polish data used as training sets, "Es" – Polish WordNet edges transferred to PWN, "Ts" – Polish texts of glosses and usage examples, languages are listed with ISO 639-1 codes. Medians and means are calculated either for 15 languages or for 6 languages (the plus sign). Statistical significance shows U-Mann-Whitney paired rank-sum test for differences between multilingual performance measures of models (in terms of F1 medians): *) $p \leq 0.05$, **) $p \leq 0.01$, ***) $p \leq 0.002$; baselines are marked with [e], [c] and [x] signs, respectively. We use arrows to mark that a tested model performs better (↑) or worse (↓) than a particular baseline, and the equal sign (=) when models are indistinguishable from baselines. The significance was corrected for false discovery ratio.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. *Small*, 8(4):5.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiaju Du, Fanchao Qi, and Maosong Sun. 2019. Using bert for word sense disambiguation. *arXiv preprint arXiv:1909.08358*.

Agnieszka Dziob, Maciej Piasecki, and Ewa Rudnicka. 2019. plWordNet 4.1 - a linguistically motivated, corpus-based bilingual resource. In *Proceedings of the 10th Global Wordnet Conference*, pages 353–362, Wroclaw, Poland. Global Wordnet Association.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense dis-

ambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907.

D. J. Kranda. 2011. The square of adjacency matrices.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411, Brussels, Belgium. Association for Computational Linguistics.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a comprehensive lexical-semantic resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2259–2268, Osaka, Japan. The COLING 2016 Organizing Committee.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

Alexander Popov. 2018. Neural network models for word sense disambiguation: an overview. *Cybernetics and information technologies*, 18(1):139–151.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A strategy of mapping polish wordnet onto princeton wordnet. In *Proceedings of COLING 2012: Posters*, pages 1039–1048.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. *arXiv preprint arXiv:2010.03017*.