Mitigating the Knowledge Acquisition Bottleneck for Hungarian Word Sense Disambiguation using Multilingual Transformers

Gábor Berend [1,2]

[1]University of Szeged, Institute of Informatics
[2]MTA-SZTE, Research Group on Artificial Intelligence
berendg@inf.u-szeged.hu

Abstract. A major hurdle in training all-words word sense disambiguation (WSD) systems for new domains and/or languages is the limited availability of sense annotated training corpora and that their construction is an extremely costly and labor-intensive process. In this paper, we investigate the utilization of multilingual transformer-based language models for performing cross-lingual WSD in the zero-shot setting. Our empirical results suggest that by relying on the intriguing multilingual abilities of pre-trained language models, we can infer reliable sense labels to Hungarian textual utterances in the all-word WSD setting by purely relying on sense-annotated training data in English.
Keywords: zero-shot word sense disambiguation; contextualized word representations; knowledge acquisition bottleneck

1 Introduction

A key difficulty in natural language understanding is definitely the highly ambiguous nature of natural language utterances. This property has made word sense disambiguation (WSD) a long-standing and central task within the NLP community (Lesk, 1986; Gale et al., 1992; Navigli, 2009) with ample application possibilities, e.g. in information retrieval (Zhong and Ng, 2012).

Most successful WSD systems are built in a monolingual and supervised manner, i.e. by having access to large amounts of sense-disambiguated training data in the same language as the test data. Obtaining such large-scale sense-annotated corpora is extremely cumbersome and known to be affected by the *knowledge acquisition bottleneck* (Gale et al., 1992), for which reason solutions that can utilize the training data composed for different languages are of upmost importance.

In this work, we evaluate WSD systems for Hungarian in the cross-lingual and zero-shot setting, as we solely use English sense-annotated data for training, whereas our primary interest is applying this model on Hungarian input texts. We bridge this potential mismatch in the language of input texts during training and test time by relying on multilingual contextualized word representations

which have been shown to yield high quality multilingual representations (Chi et al., 2020; Dufter and Schütze, 2020) that makes them suitable for application in cross-lingual zero-shot settings. We make our source code for reproducing our experiments available at `https://github.com/begab/sparsity_makes_sense`.

2 Related work

Contextualized word representations, such as CoVE (McCann et al., 2017), ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), are the most prominent forms of representing the meaning of linguistic units nowadays. Contextualized word representations are *typically* built on the transformer architecture (Vaswani et al., 2017) by employing some kind of masked language modeling objective.

The fact that these models can be trained in a self-supervised manner, i.e. they do not require any explicit labeled data, but raw text only, allows them to be trained on data at an unprecedented scale. As a consequence of being trained on a wide variety of textual utterances, such models have the ability of developing such representations that can capture a wide variety of linguistic phenomena Tenney et al. (2019); Hewitt and Manning (2019); Reif et al. (2019). This is a useful byproduct of these architectures as their training procedures do not explicitly encourage them in becoming able to capture these linguistic phenomena.

Transformer-based language models trained on multilingual texts – without supposing any kind of alignment between the multilingual text passages – have been shown to provide such representations that perform surprisingly well across different languages (K et al., 2020; Chi et al., 2020; Dufter and Schütze, 2020). This property of multilingual transformers opens the possibility of utilizing them in zero-shot settings where the language of a training data do not need to match that of the test set for modeling certain linguistic phenomena.

Recent studies have shown that transformer-based language models that provide contextualized word representations can output extremely valuable inputs to WSD systems even when they are applied in a simple k-nearest neighbor classifier (Loureiro and Jorge, 2019). The application of contextualized word representations with additional sparsity constraints have been recently reported to yield extra improvement for WSD (Berend, 2020a).

These earlier works were, however, focusing on the evaluation of applying contextualized word representations in monolingual WSD settings, i.e. when the sense annotated training corpus is available in the same language relative to the test set. Our work is different from this prior line of research in that we are investigating the performance of these techniques when used in conjunction with multilingual contextualized word representations and evaluate them in the zero-shot setting when the language of the sense annotated training set does not necessarily match the language of the test set.

Most recently, Scarlini et al. (2020) proposed ARES (context-AwaRe Embeddings of Senses) for obtaining sense prototype embeddings that can be used

for WSD in a 1-nearest neighbor fashion similar to (Loureiro and Jorge, 2019). ARES introduces a methodology to exploit useful signal upon the construction of sense embeddings from external knowledge stored in the SyntagNet (Maru et al., 2019). ARES embeddings for all the English WordNet synsets obtained by relying on multilingual BERT were made publicly available by the authors, which makes their utilization possible for languages beside English as well.

3 Experiments

Most successful WSD systems are based on supervision (Raganato et al., 2017). That is, they require (large) sense-annotated training signal in the language of the test sentences. The largest sense annotated training data are in English and use the Princeton WordNet (Fellbaum, 1998) as the basis of sense inventory for disambiguating the distinct senses of the words in their context.

Our evaluation departs from the typical setting, i.e. we rely on English sense-annotated training data for training and evaluate the created WSD model by distinguishing senses of ambiguous words in Hungarian sentences.

3.1 Dataset

We first introduce the English sense-annotated corpora that we used for training our WSD models. We also provide details on the evaluation datasets in both English and Hungarian that we evaluated our models on.

English data We evaluated our approach using the unified WSD evaluation framework (Raganato et al., 2017) that includes the sense-annotated SemCor dataset for training purposes. SemCor Miller et al. (1994) contains 802,443 tokens, out of which more than 28% (226,036) is sense-annotated according to WordNet sensekeys.

We also used the contents of the glosses of the English WordNet synsets and the Princeton WordNet Gloss Corpus (WNGC) as additional sources for constructing our models. WNGC is a sense-annotated version of the WordNet definitions themselves, hence it can be basically used as an extension to the SemCor annotated training set (Vial et al., 2019).

The evaluation framework introduced in (Raganato et al., 2017) also contains five different all-words WSD benchmarks for measuring the performance of WSD systems in English. We also used those for measuring the performance of our WSD models that were based on the contextualized word representations from a multilingual language model. This means that the self-supervised pre-training was conducted over multilingual data, but the creation of our all-words WSD model was both trained and evaluated on English for these experiments. Our evaluation dataset in English consisted of the concatenated test set of the SensEval2 Edmonds and Cotton (2001), SensEval3 Mihalcea et al. (2004), SemEval 2007 Task 17 Pradhan et al. (2007), SemEval 2013 Task 12 Navigli et al.

(2013), SemEval 2015 Task 13 Moro and Navigli (2015) shared tasks, comprising of 7253 sense-annotated tokens in total. During our evaluation, we relied on the official scoring script included in the evaluation framework from (Raganato et al., 2017) in our monolingual experiments.

Hungarian test set The dataset we used for our experiments is from (Berend, 2020b), which is a distilled version of the sense-annotated corpus introduced in Vincze et al. (2008). The original dataset contains a collection of documents written in Hungarian that are part of the Hungarian National Corpus (HNC) (Váradi, 2002). The difference between the two datasets is that whereas the former consists of sentences containing ambiguous words, the latter also contains the entire documents for those sentences.

Our corpus that we used for evaluating the performance of our WSD models in Hungarian includes 12,477 sentences, each containing an ambiguous word along with its sense ambiguated label. The sense-annotated dataset contains sense-disambiguated occurrences of 39 different word forms (and their morphologically inflected variants). The 39 distinct word forms were manually disambiguated to one of 200 distinct senses.

3.2 Approach

It was shown recently that by using transformer-based masked language models, such as BERT (Devlin et al., 2019), it becomes possible to build WSD systems by simply obtaining the contextualized embeddings for occurrences of ambiguous words from a sense annotated corpora and performing WSD using nearest neighbor classification for test words (Loureiro and Jorge, 2019). This approach (coined as LMMS by its authors) was, however, both solely trained and evaluated on the unified WSD framework from Raganato et al. (2017), in which cross-linguality did not play a role as both the training and validation corpora were in English.

A recent modification of the LMMS approach proposed the utilization of sparse contextualized word representations and the reliance on the analysis of the sparsity structure of the sense annotated word vectors (Berend, 2020a). We shall refer to this variant of the LMMS approach as S-LMMS throughout the paper, where the prefix is meant to denote that this model variant is based on sparse contextualized word representations.

Performing WSD using the S-LMMS algorithm has been reported to outperform the LMMS strategy significantly when being both trained and evaluated on English, however, it remains a question if the superiority of S-LMMS can be observed in the cross-lingual unsupervised setting as well. Additionally, we explore the effects of using different transformer-based masked language models in our experiments, whereas the original work reported results only on the application of the large cased BERT model.

S-LMMS has two hyperparameters, the dimensionality of the sparse vectors K and the regularization coefficient λ which controls the level of sparsity of the

vectors. We decided to use the same values that was suggested in the original paper, i.e. $K = 3000$ and $\lambda = 0.05$.

3.3 Evaluation

Our evaluation ranges over the investigation of multiple transformer-based multilingual masked language models as inputs to the algorithms we conducted our experiments with, i.e. we relied on the application of multilingual BERT (Devlin et al., 2019), referenced as mBERT hereon, and different versions of the multilingual XLM-Roberta (Conneau et al., 2020) architecture, the XLM-Roberta base and large models.

We used the transformers library (Wolf et al., 2020) to obtain the contextualized multilingual embeddings for our experiments. Even though there exist BERT models specially dedicated to the processing of Hungarian texts, e.g. (Nemeskey, 2020), applying such monolingual models would not suit our setting, since there is a shortage of sense-annotated training data of reasonable size for Hungarian that would be required for training the WSD models. Hence, we mitigate the knowledge acquisition bottleneck of obtaining a high-coverage all-words WSD dataset in Hungarian by using multilingual transformer models and sense-annotated training data in English.

For each model we experiment with, we evaluate the utility of using the contextualized embeddings from the last four layers of the transformer models as well as taking the average of these four layers. The base models consist of 12 layers, whereas the large transformer model has 24 layers.

3.4 Evaluation in English

First, we trained the LMMS and S-LMMS all words WSD models in English and evaluated their utility in the monolingual setting, i.e. on the standard benchmark validation set from (Raganato et al., 2017). Table 1 contains the F1-scores of the different models when relying on contextualized embeddings produced by different multilingual transformer architectures.

Interestingly, the performance of the LMMS models are relatively stable when using multilingual contextualized transformer models of different size as input, however, the Roberta large model has a clear advantage once the contextualized representations it produces are fed into the S-LMMS approach.

Table 1 additionally confirms the findings from (Berend, 2020a), i.e. the S-LMMS variant has a clear advantage over the application of the LMMS model for all the layers and transformer models. There is nonetheless one key methodological difference in our experimental settings compared to those of (Berend, 2020a). Even though the S-LMMS approach and the training/validation data were the same, our work builds on top of multilingual contextualized word representations as input, instead of using the English monolingual BERT large model as input.

The reliance on multilingual models caused some performance loss compared to the best results reported in (Berend, 2020a) (75.7 vs. 78.8), nonetheless the

Layer(s)	LMMS	S-LMMS
9	0.698	**0.740**
10	0.702	**0.742**
11	0.706	**0.746**
12	0.699	**0.738**
last-4	0.704	**0.743**

(a) mBERT

Layer(s)	LMMS	S-LMMS
9	0.684	**0.735**
10	0.701	**0.749**
11	0.702	**0.753**
12	0.686	**0.736**
last-4	0.695	**0.748**

(b) XLM-Roberta-base

Layer(s)	LMMS	S-LMMS
21	0.702	**0.757**
22	0.692	**0.753**
23	0.679	**0.749**
24	0.648	**0.728**
last-4	0.692	**0.754**

(c) XLM-Roberta-large

Table 1. F-scores obtained over the standard English WSD evaluation benchmark dataset from (Raganato et al., 2017). Results range over the application of various final layers (and their averaging) of different transformer architectures as input to LMMS and S-LMMS.

application of multilingual contextualized transformation is a key component for using the trained model for all-words WSD purposes in the cross-lingual setting.

3.5 Evaluation in Hungarian

Subsequently, we conducted experiments for assessing the quality of the WSD models when applied on Hungarian input texts. When evaluating our models in this cross-lingual setting, we faced the problem that the word senses our models produce are based on the sense inventory of the English WordNet, however, there is no one-to-one correspondence between the synsets of the English WordNet and the senses of the Hungarian sense-disambiguated corpus we used for evaluation.

Our first attempt was create a manual alignment between the senses in the Hungarian dataset and the English WordNet, however, no one-to-one correspondence could be established between the two inventories, as the Hungarian dataset includes such senses that either do not exist in the WordNet or which can correspond to multiple WordNet sysnets.

To overcome this issue, we first decide on an evaluation metric that was originally introduced for measuring the performance of clustering techniques. This metric is the V-score (Rosenberg and Hirschberg, 2007), which is similar in nature to the well-known F-score that is intended to be used for the evaluation of classification algorithms. Applying V-score is advantageous, as it can handle situations when we the number of predicted categories and that of the gold standard labels mismatch. This was exactly the situation during our evaluations as we assigned one of the 117,659 English WordNet synsets to the target words that were labeled according to one of the 200 gold standard labels in the dataset. As such, we had 200 gold standard groups of words on the one hand, and as many predicted clusters as many distinct WordNet synsets were assigned to the ambiguous words in the dataset by our algorithms on the other hand.

Just like F-score is the harmonic mean of the precision and the recall of an algorithm, V-score is the harmonic mean of the homogeneity and the complete-

ness scores that are meant to be a respective generalization of the precision and recall scores employed for clustering.

The V-scores between the predicted synset labels from the English WordNet and the gold standard senses according to our evaluation dataset over the 12,477 sense-disambiguated Hungarian words are included in Table 2.

Even though applying V-score for assessing the quality of the implicit clustering based on the most likely synset our models predicted for the sense-annotated Hungarian words is a viable approach, it arguably gives us a too pessimistic view on the true quality of our models. This is partly because our model predicts one of the 117,659 different English WordNet synsets, whereas there were only 200 distinct senses distinguished in the sense-annotated corpus we based our evaluation on.

As such, our model often ended up producing near-misses, such as the (co-)hypernym/hyponym of the correct senses of a word. Additionally, due to the mismatch of the employed labels, there were certain cases when there would be no exact match in the English WordNet for some Hungarian sense labels.

Layer(s)	LMMS	S-LMMS
9	**0.216**	0.209
10	0.218	**0.220**
11	0.197	**0.219**
12	0.196	**0.207**
last-4	**0.220**	**0.220**

(a) mBERT

Layer(s)	LMMS	S-LMMS
9	**0.242**	0.236
10	0.259	**0.261**
11	0.255	**0.262**
12	**0.239**	0.233
last-4	0.254	**0.258**

(b) XLM-Roberta-base

Layer(s)	LMMS	S-LMMS
21	0.262	**0.277**
22	0.246	**0.272**
23	0.229	**0.262**
24	0.224	**0.227**
last-4	0.253	**0.274**

(c) XLM-Roberta-large

Table 2. V-scores averaged over the ambiguous word forms obtained from the Hungarian WSD evaluation benchmark dataset. Results range over the application of various final layers (and their averaging) of different transformer architectures as input to LMMS and S-LMMS.

Due to the above reasons, we assessed the quality of our cross-lingual WSD predictions in an alternative manner for our subsequent experiment. Instead of expecting the models to find a good one-to-one mapping between the English synsets and the set of sense labels included in our Hungarian evaluation set (which does not even exist in the first place for certain senses by the design of the sense labels of the two different sense inventories), we quantified the extent to which the ordered list of the English synsets that our models assigned to the Hungarian ambiguous words are similar for those words that received the same sense label in our evaluation dataset in Hungarian.

For each ambiguous Hungarian word in our test set, we determined the top-15 English synsets our model assigned to them. As a subsequent step, we calculated the similarity between all pairs of ambiguous words based on the ordered list of most likely English synsets assigned to the word occurrences.

Finally, we determined the nearest neighbor of each ambiguous word according to their similarity and quantified the relative number of times the ground truth sense label of nearest neighbor words were identical. As such, these results can be viewed as the performance of a 1-nearest-neighbor classifier that determines the proximity of word occurrences based on the ordered list of most likely synsets that our cross-lingual model assigns to them.

The top-15 most likely synsets assigned to a pair of ambiguous words are non-conjoint, meaning that they can differ to any extent. Indeed, in the most extreme case, there could be no overlap at all between the ranked lists of synsets. To this end, we measured the similarities between the top-ranked synsets for a pair of words using the pairwise ranking-biased overlap (RBO) (Webber et al., 2010) score, which (among others) has the favorable property of being capable of measuring the similarity between non-conjoint ordered lists.

These results are included in Table 3. Similar to the earlier results in Table 1 and Table 2, i.e. the utilization of S-LMMS with input representations originating from the 21th layer of the large XLM-Roberta model provided the best results according to the nearest-neighbor accuracy metric.

Layer(s)	LMMS	S-LMMS
9	**0.756**	0.740
10	**0.756**	0.752
11	0.740	**0.742**
12	**0.730**	0.716
last-4	**0.761**	0.757

(a) mBERT

Layer(s)	LMMS	S-LMMS
9	0.792	**0.795**
10	0.811	**0.815**
11	0.808	**0.815**
12	**0.796**	0.795
last-4	0.801	**0.805**

(b) XLM-Roberta-base

Layer(s)	LMMS	S-LMMS
21	0.828	**0.830**
22	0.818	**0.821**
23	0.809	**0.814**
24	**0.791**	0.777
last-4	0.823	**0.824**

(c) XLM-Roberta-large

Table 3. Nearest neighbor accuracy averaged over the ambiguous word forms obtained from the Hungarian WSD evaluation benchmark dataset. Results range over the application of various final layers (and their averaging) of different transformer architectures as input to LMMS and S-LMMS.

3.6 Comparison with ARES embeddings

As mentioned in Section 2, ARES embeddings (Scarlini et al., 2020) can be utilized for WSD similar to LMMS (Loureiro and Jorge, 2019) by performing a 1-nearest neighbor search between the contextualized embedding of some word and the pre-calculated sense embeddings. The important difference between ARES and LMMS is that ARES also uses a semi-supervised approach to improve the sense embeddings obtained for those WordNet synsets with no/few occurrences in the sense-annotated training corpus, e.g. SemCor. A key important property of the ARES embeddings from our point is that such a variant that builds upon the contextualized representations of the last 4 layers of mBERT is made available.

In Table 4, we compare our previously reported results to those obtained with ARES. For better comparability, we chose to compare the results of the (S-)LMMS approaches that rely on the same contextualized representations as ARES, i.e. the averaged outputs of the last 4 layers of mBERT. S-LMMS performed clearly better on the English test set, whereas our evaluation on the Hungarian data shows a mixed, but comparable behavior for the different approaches.

	ARES	LMMS	S-LMMS
Accuracy	0.713	0.704	**0.743**

(a) Evluation on English.

	ARES	LMMS	S-LMMS
V-score	0.208	**0.220**	**0.220**
Accuracy	**0.765**	0.761	0.757

(b) Evaluation on Hungarian.

Table 4. Comparing the application of ARES embeddings and our models using the last 4 layers of mBERT embeddings.

4 Conclusions

Transformer-based language models are known to provide extremely valuable input for WSD. Our paper investigated the possibility of exploiting the sense-annotated training corpus of some resource-rich source language, and utilizing the trained WSD model on some distinct target language by relying on the cross-lingual interoperability of the multilingual contextualized word representation produced by mBERT and XLM-Roberta. The proposed approach showed a promising solution for overcoming the need of a large high-coverage sense-annotated training corpus.

Acknowledgments

The research was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program and the Artificial Intelligence National Excellence Program (grant no.: 2018-1.2.1-NKP-2018-00008).

The author is grateful for the fruitful discussions on the topic of word sense disambiguation with Ádám Szórád whose research was funded by the project "Integrated program for training new generation of scientists in the fields of computer science", no EFOP-3.6.3-VEKOP-16-2017-0002, supported by the EU and co-funded by the European Social Fund.

Bibliography

Berend, G.: Sparsity makes sense: Word sense disambiguation using sparse contextualized word representations. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8498–8508. Association for Computational Linguistics, Online (Nov 2020a), https://www.aclweb.org/anthology/2020.emnlp-main.683

Berend, G.: Word sense disambiguation for hungarian using transformers. In: Berend, G., Gosztolya, G., Vincze, V. (eds.) XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020). p. 3–13. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2020b)

Chi, E.A., Hewitt, J., Manning, C.D.: Finding universal grammatical relations in multilingual BERT. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5564–5577. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.acl-main.493

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.acl-main.747

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://www.aclweb.org/anthology/N19-1423

Dufter, P., Schütze, H.: Identifying elements essential for BERT's multilinguality. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4423–4437. Association for Computational Linguistics, Online (Nov 2020), https://www.aclweb.org/anthology/2020.emnlp-main.358

Edmonds, P., Cotton, S.: SENSEVAL-2: Overview. In: The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems. pp. 1–5. SENSEVAL '01, Association for Computational Linguistics, Stroudsburg, PA, USA (2001), http://dl.acm.org/citation.cfm?id=2387364.2387365

Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)

Gale, W.A., Church, K.W., Yarowsky, D.: A method for disambiguating word senses in a large corpus. Computers and the Humanities 26(5), 415–439 (Dec 1992), https://doi.org/10.1007/BF00136984

Hewitt, J., Manning, C.D.: A structural probe for finding syntax in word representations. In: Proceedings of the 2019 Conference of the North Ameri-

can Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4129–4138. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://www.aclweb.org/anthology/N19-1419

K, K., Wang, Z., Mayhew, S., Roth, D.: Cross-lingual ability of multilingual BERT: An empirical study. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=HJeT3yrtDr

Lesk, M.: 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. pp. 24–26. SIGDOC '86, ACM, New York, NY, USA (1986), http://doi.acm.org/10.1145/318723.318728

Loureiro, D., Jorge, A.: Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5682–5691. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://www.aclweb.org/anthology/P19-1569

Maru, M., Scozzafava, F., Martelli, F., Navigli, R.: SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. 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). pp. 3534–3540. Association for Computational Linguistics, Hong Kong, China (Nov 2019), https://www.aclweb.org/anthology/D19-1359

McCann, B., Bradbury, J., Xiong, C., Socher, R.: Learned in translation: Contextualized word vectors. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 6294–6305. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7209-learned-in-translation-contextualized-word-vectors.pdf

Mihalcea, R., Chklovski, T., Kilgarriff, A.: The senseval-3 english lexical sample task. In: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. pp. 25–28. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://www.aclweb.org/anthology/W04-0807

Miller, G.A., Chodorow, M., Landes, S., Leacock, C., Thomas, R.G.: Using a semantic concordance for sense identification. In: HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 (1994), https://www.aclweb.org/anthology/H94-1046

Moro, A., Navigli, R.: SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 288–297. Association for Computational Linguistics, Denver, Colorado (Jun 2015), https://www.aclweb.org/anthology/S15-2049

Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41(2) (Feb 2009), https://doi.org/10.1145/1459352.1459355

Navigli, R., Jurgens, D., Vannella, D.: SemEval-2013 task 12: Multilingual word sense disambiguation. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 222–231. Association for Computational Linguistics, Atlanta, Georgia, USA (Jun 2013), https://www.aclweb.org/anthology/S13-2040

Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020)

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://www.aclweb.org/anthology/N18-1202

Pradhan, S.S., Loper, E., Dligach, D., Palmer, M.: Semeval-2007 task 17: English lexical sample, srl and all words. In: Proceedings of the 4th International Workshop on Semantic Evaluations. pp. 87–92. SemEval '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007), http://dl.acm.org/citation.cfm?id=1621474.1621490

Raganato, A., Camacho-Collados, J., Navigli, R.: Word sense disambiguation: A unified evaluation framework and empirical comparison. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 99–110. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://www.aclweb.org/anthology/E17-1010

Reif, E., Yuan, A., Wattenberg, M., Viegas, F.B., Coenen, A., Pearce, A., Kim, B.: Visualizing and measuring the geometry of bert. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32, pp. 8594–8603. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf

Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). pp. 410–420. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https://www.aclweb.org/anthology/D07-1043

Scarlini, B., Pasini, T., Navigli, R.: With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 3528–3539. Association for Computational Linguistics, Online (Nov 2020), https://www.aclweb.org/anthology/2020.emnlp-main.285

Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computa-

tional Linguistics. pp. 4593–4601. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://www.aclweb.org/anthology/P19-1452

Váradi, T.: The Hungarian national corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain (May 2002), http://www.lrec-conf.org/proceedings/lrec2002/pdf/217.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30, pp. 5998–6008. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vial, L., Lecouteux, B., Schwab, D.: Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In: Global Wordnet Conference. Wroclaw, Poland (2019), https://hal.archives-ouvertes.fr/hal-02131872

Vincze, V., Szarvas, Gy., Almási, A., Szauter, D., Ormándi, R., Farkas, R., Hatvani, Cs., Csirik, J.: Hungarian word-sense disambiguated corpus. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (May 2008)

Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. 28(4) (Nov 2010), https://doi.org/10.1145/1852102.1852106

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6

Zhong, Z., Ng, H.T.: Word sense disambiguation improves information retrieval. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 273–282. Association for Computational Linguistics, Jeju Island, Korea (Jul 2012), https://www.aclweb.org/anthology/P12-1029