Word Sense Disambiguation for Hungarian using 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. In this paper we investigate the applicability of contextual word embeddings for the task of word sense disambiguation (WSD) in Hungarian. We show that a simple k–nn (k–nearest neighbors) approach which relies on multilingual BERT representations can yield highly accurate results in terms of F-scores when evaluated for word sense disambiguation.

Keywords: contextual word representations; multilingual BERT; word sense disambiguation (WSD)

1 Introduction

Word embeddings have been prevalently applied in a variety of natural language processing applications ranging from machine translation (Bahdanau et al., 2014) to information retrieval (Vulić and Moens, 2015) and sentiment analysis (Socher et al., 2013), among others.

A major shortcoming of standard static word embeddings, including `word2vec` (Mikolov et al., 2013) and Glove (Pennington et al., 2014) is that they assign a fixed representation to the individual word forms. That is, the vectorial representations belonging to a word is fixed and it behaves agnostically to the context a particular word is presented. Until recently, such word representations have dominated NLP applications.

Contextualized word representations, such as CoVe (McCann et al., 2017), ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), however, have the added favorable property that they are capable of incorporating the context in which a particular word is mentioned upon constructing its vectorial representation. This characteristic of contextualized word embeddings makes them highly appealing for applying them to the task of word sense disambiguation (WSD), where the task is to choose the most appropriate sense a particular word form has based on its context.

There have been some investigation of applying contextual word embeddings for WSD in English (Loureiro and Jorge, 2019; Vial et al., 2019). Our paper is complementary to these results in that here we give a thorough empirical evaluation for using contextual word embeddings for performing WSD in Hungarian.

Our solution uses a simple, yet effective k–nn-based approach for performing WSD. The main contributions of the paper are that

- we evaluate and carefully analyze the applicability of the off-the-shelf multilingual BERT model being applied for Hungarian WSD by a k–nn based approach,
- make the contextualized word embeddings obtained for nearly 12500 sense-annotated utterances publicly available.

2 Related work

One of the key difficulties of natural language understanding is the highly ambiguous nature of language. As a consequence, WSD has long-standing origins in the NLP community Lesk (1986) and it is still in the focus of a series of recent research efforts in NLP (Raganato et al., 2017; Melamud et al., 2016; Loureiro and Jorge, 2019; Vial et al., 2019).

The typical setting for WSD is to categorize the mentions of ambiguous words according to some sense inventory. The most frequently applied sense inventory in the case of English is definitely the Princeton WordNet (Fellbaum, 1998). A Hungarian version of the WordNet also has been created (Miháltz et al., 2008) serving the basis of the Hungarian WSD dataset created by Vincze et al. (2008).

WSD systems either take some unsupervised, knowledge-based or some supervised approach requiring a training corpus with sense-annotated utterances of ambiguous words. Unsupervised approaches could attempt to match the mentions of ambiguous words to their proper sense based on the textual overlap between the context of an ambiguous word and the definitions included to its potential senses according to the sense inventory employed (Lesk, 1986) or be based on random walks over the semantic graph providing the sense inventory (Agirre and Soroa, 2009).

Supervised WSD techniques typically perform better than unsupervised approaches. IMS (Zhong and Ng, 2010) is a classical supervised WSD framework which was created with the intention of easy extensibility. It uses an SVM classifier, which derives features for an ambiguous word based on the word forms and POS tags of the words in its neighborhood. The recent advent of neural text representations have also shaped the landscape of algorithms performing WSD. Melamud et al. (2016) devised the context2vec framework, which relies on a bidirectional LSTM for performing supervised WSD. Most recently, (Loureiro and Jorge, 2019; Vial et al., 2019) have proposed the usage of contextualized word representations for tackling WSD.

Contextualized word representations (McCann et al., 2017; Peters et al., 2018; Devlin et al., 2019) are recent extensions of traditional word embeddings, such as `word2vec` (Mikolov et al., 2013), with the notable distinction that they construct different vectorial representation even for the same word form when employed in a different context. Contextualized word representations employ some language modeling inspired objective and are trained on massive amounts of textual data, which makes them generally applicable in a variety of settings, including natural language inference (Williams et al., 2018) or reading comprehension (Khashabi et al., 2018).

3 Experiments

We next introduce the dataset we performed our experiments on, as well as the kind of contextual word representations we determined for it.

3.1 The dataset

The dataset we performed our experiments on is derived from the sense-annotated corpus introduced by Vincze et al. (2008). The dataset contains a collection of documents written in Hungarian that are part of the Hungarian National Corpus (HNC) (Váradi, 2002) including mentions towards 39 ambiguous words. The documents are selected from the *Heti Világgazdaság* subcorpus containing mostly news documents related to business and politics. The different word senses got disambiguated in compliance with the sense inventory of the Hungarian WordNet (Miháltz et al., 2008).

The corpus released by Vincze et al. (2008) contains the entire documents in which the sense-annotated ambiguous words are located. The original dataset contains a separate file for each of the word forms in an ISO-8859-1 encoded XML file. We distilled the original WSD corpus (Vincze et al., 2008) into a single and easy-to-handle tab-separated plain text file in UTF-8 format. The distilled version of the dataset differs from the original dataset in that it contains only the local context of the ambiguous words as opposed to the entire document they are included in. We make this dataset accessible [1], a sample line from which is

```
anyagi_a_1_pénzzel_kapcsolatos 1 Az anyagi kár meghaladja az egymilliárd
schillinget .
```

with the first string denoting the ground truth sense label for the ambiguous word, the second item in the line denoting the token position of the ambiguous target word within the excerpt, followed by the excerpt itself in a tokenized format. The entire dataset contains 12477 distinct mentions for one of the 39 ambiguous Hungarian words. The 12477 excerpts contain a total of 449875 tokens.

Figure 1 illustrates the joint distribution of the number of senses per word forms and the Shannon entropy quantifying the heterogeneity of the distributions of the different senses of word forms. We can see that the number of word senses listed for a particular word form ranged between 1 (for the word form *tanár*/teacher) and 14 (for the word form *jár*/go). Perhaps unsurprisingly, a

[1] http://github.com/begab/huWSDdata

strong positive correlation of $\rho = 0.83$ can be observed between the two quantities, i.e. the higher number of distinct meanings a word form has, the higher amount of uncertainty can be observed on average regarding the predictability of its actual meaning in context.

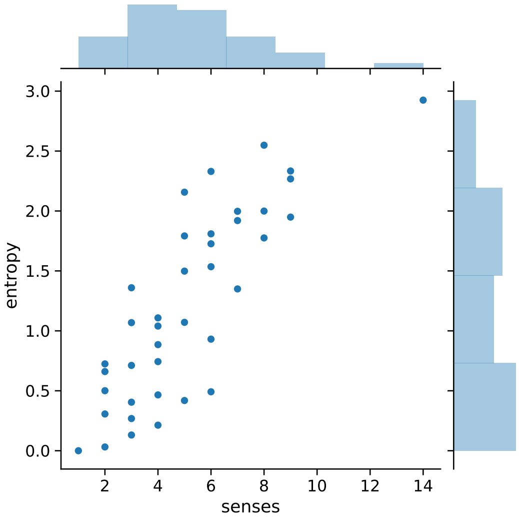


Fig. 1. The joint distribution of the number of distinct senses and the Shannon entropy of their distributions for the 39 word forms in the Hungarian WSD dataset.

3.2 Preprocessing the dataset

We preprocessed the previously introduced WSD dataset using the pretrained cased multilingual BERT (M-BERT) architecture for obtaining contextual word representations. This preprocessing step was conducted using the `Huggingface transformers` Python package (Wolf et al., 2019). We defined the contextualized vectorial form of the individual tokens in the excerpts as the average of the vectorial representations of the word pieces as determined by the M-BERT cased multilingual tokenizer.

The pretrained M-BERT model uses a transformer model which has one word piece-based input layer, followed by 12 stacked layers using self-attention. Each of the 12+1 layers are identical in that they employ vectorial representations of 768 dimensions. We calculated and evaluated the 768-dimensional contextualized word representations for every token. We also performed a sensitivity analysis on using the contextual word representations originating from the different layers of the multi-layered transformer model of M-BERT (cf. Figure 2).

We managed to determine contextual word representations for all but one of the 12477 sense-annotated words in our dataset. The reason why we had to omit one of the sense-annotated words from our analysis was that it was included in an excerpt being longer than the longest sequence M-BERT architecture can possibly deal with, i.e. a sequence length of 512. We also release our contextualized embeddings for the 12476 sense-annotated words that we determined M-BERT representations for at http://github.com/begab/huWSDdata.

3.3 Results

We first review the results obtained in (Vincze et al., 2008) using a traditional approach that is similar to the one applied in IMS (Zhong and Ng, 2010). We subsequently introduce our approach for performing WSD using contextualized M-BERT representations and report our quantitative results.

Overview of the findings from (Vincze et al., 2008) Similar to how it was done in our experiments, Vincze et al. (2008) relied only on the context to be found in the local proximity of the sense disambiguated word forms. The ambiguous words were then represented using the traditional vector space model (VSM) based on the context in the same paragraph of sense-annotated ambiguous words. The features determined for an ambiguous token could additionally include indicator features based on the directly surrounding 3 words of some target word. Vincze et al. (2008) also made use of the POS tag information of the tokens, i.e. they considered only the lemmatized word forms of nouns, verbs, adjectives and adverbs as contextual features from the vicinity of a target token for constructing their feature vector.

Based on the above representation of sense-annotated word forms, Vincze et al. (2008) reports a micro-averaged F-score of 0.703 when relying on a Naïve Bayes classifier and evaluation metrics ranging between 0.727 and 0.749 for applying C4.5 classifier depending on the combination of features they were relying on. Vincze et al. (2008) used a leave-one-out evaluation for assessing the quality of their classifiers for performing WSD. That is, each time a new model was trained on all but one of the feature vectors belonging to the different senses of one of the ambiguous word forms and evaluation was performed against the single one ambiguous instance that was held out from the training instances.

(Vincze et al., 2008) reported evaluation scores for the simple – but often difficult to beat – baseline for always predicting the Most Frequent Sense (MFS) of an ambiguous word, regardless of its context. The MSF baseline obtained an aggregate micro-averaged F-score of 0.694.

Using contextual representations for WSD Our methodology for applying M-BERT representations to WSD is similar to those recently proposed in (Loureiro and Jorge, 2019) for English WSD. An important technical difference between (Loureiro and Jorge, 2019) and our work is that while (Loureiro and Jorge, 2019) based their experiments on the large cased BERT model dedicated to the English language alone, we were utilizing the multilingual BERT (M-BERT) model in order to be able to use it for WSD in Hungarian. Note that we did not perform any fine-tuning of the M-BERT model to fit the task of WSD, but simply used the pre-trained model in our approach.

The way we evaluated the utilizability of M-BERT embeddings for inclusion in word sense disambiguating the utterances of ambiguous words in Hungarian was via integrating it in a simple k–nn classifier based on the contextualized word vectors determined for the sense-annotated tokens. That is, for a pre-defined value of k and some query word q along with its contextualized word vector $\mathbf{q}$, we simply looked for its k closest neighbor among the sense-annotated contextualized word vectors and returned the majority vote for the sense annotations of the training instances according to their ground truth senses. Similar to (Vincze et al., 2008), we also conducted experiments in a leave-one-out fashion.

We repeated our experiments when relying on different number of nearest neighbors, i.e. $k \in \{1, 3, 5, 7, 9\}$. Figure 2 illustrates the effect of choosing the value for k differently when relying on the M-BERT representation originating from the different layers of the transformer architecture. Figure 2 corroborates previous results on contextual representations that the topmost layers tend to perform better in general, especially for evaluations related to semantics. Results reported in Figure 2 also show a plateauing effect for the last few layers of M-BERT contextualized embeddings. That is, no great improvements can be witnessed when utilizing M-BERT representations derived from the layers in the range of 8 to 12. The earlier layers, however, performed subpar to the final layers.

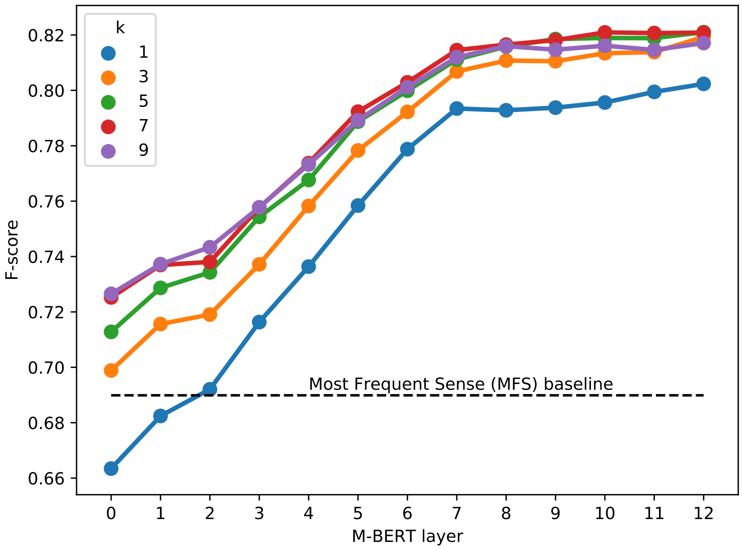
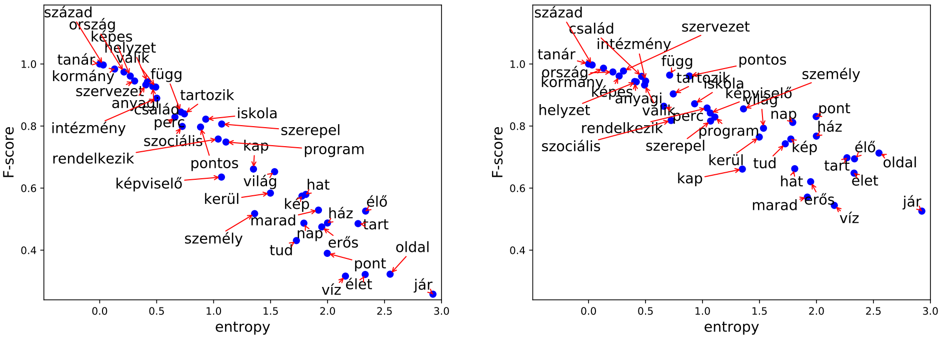


Fig. 2. Aggregated results over the 39 word forms for the MSF baseline and the k–nn model based on M-BERT, when using different values for k.

Figure 2 also shows that increasing the value for the nearest neighbors considered in the prediction can improve performance. Setting k too high, however, is not a good idea, since that would hamper the identifiability of rare senses, and the identification of uncommon senses could often be of potential interest. Hence we argue that using the median from the tested values for k, i.e. $k = 5$, provides a trade-off between delivering increased performance – as opposed to choosing smaller values of k – and being less biased in predicting (the most) frequent senses – as opposed to applying higher values of k.

We can also see it in Figure 2 that k-nn models based on the M-BERT contextual word representations obtained from layer 5 and beyond are outperforming the best reported results in (Vincze et al., 2008) irrespective of the value of k employed. Note that when relying on the final layers of the transformer architecture and employing $k > 3$, we consistently managed to outperform the best previous results by a fair margin (cf. 0.74 versus 0.82).



(a) MFS baseline

(b) 5–nn using the last M-BERT layer

Fig. 3. The Shannon entropy of the word sense distributions and the aggregated F-scores of the senses for the individual word forms in the dataset.

As a final assessment, we compared the performance of the MFS baseline and our k–nn solution relying on the M-BERT contextualized representations on the individual level of ambiguous word forms. This comparison contrasts the Shannon entropy of the sense distribution an ambiguous word form has and the F-score obtained for it for a particular model. These results are included in Figure 3 for the MFS and the 5–nn approach relying on the final layer of M-BERT representations for disambiguation.

We can see that while the performance of the MFS baseline fluctuates heavily – with 5 out of 39 word forms having an F-score less than 0.4 – the 5–nn model manages to deliver an F-score at least 0.577, even for the most ambiguous word form (*jár*/go).

We calculated the Person correlation between the results reported in Figure 3. The Shannon-entropy for the sense distribution a word form has and the performance the different models can achieve for them come hand in hand with a strong negative correlation between the two values. For the MFS and the 5–nn approaches reported in Figure 3 we observed Pearson correlation coefficients of -0.968 and -0.896, respectively. The mere fact that it is more difficult to predict the proper sense for words with a more diverse set of meanings (hence a higher Shannon-entropy) is not so surprising. It would be nonetheless interesting to investigate the reasons for the k–nn based approach behaving less sensitively to the diversity of the ambiguous word forms.

4 Future work and conclusions

This paper focused on WSD in Hungarian for a relatively small set of 39 specific word forms. In order to increase the real world applicability of our model, we plan to extend it to the more challenging all words WSD setting. Training datasets annotated for the all words WSD problem are available in English Raganato et al. (2017); Taghipour and Ng (2015), however, such large scale training data is not currently available for Hungarian at the moment. As a future research, our goal is to investigate how already existing sense-annotated training data – in some possibly foreign language – can improve the performance of WSD.

In this paper, we investigated the extent to which multilingual BERT provides a useful representation for word sense disambiguation. We have seen that a simple solution which uses a k–nn approach for determining the sense of an ambiguous word based on its contextual word representation can obtain highly accurate results.

Acknowledgement

This research was partly 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. This work was in part supported by the National Research, Development and Innovation Office of Hungary through the Artificial Intelligence National Excellence Program (grant no.: 2018-1.2.1-NKP-2018-00008).

Bibliography

Agirre, E., Soroa, A.: Personalizing pagerank for word sense disambiguation. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. pp. 33–41. EACL ’09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1609067.1609070

Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014), http://arxiv.org/abs/1409.0473, cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation

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

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

Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., Roth, D.: Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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. 252–262. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://www.aclweb.org/anthology/N18-1023

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

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

Melamud, O., Goldberger, J., Dagan, I.: context2vec: Learning generic context embedding with bidirectional LSTM. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. pp. 51–61. Association for Computational Linguistics, Berlin, Germany (Aug 2016), https://www.aclweb.org/anthology/K16-1006

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the hungarian wordnet project. In: Proceedings of The Fourth Global WordNet Conference. pp. 311–321 (2008)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013)

Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in

Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014), https://www.aclweb.org/anthology/D14-1162

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

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

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013), https://www.aclweb.org/anthology/D13-1170

Taghipour, K., Ng, H.T.: One million sense-tagged instances for word sense disambiguation and induction. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning. pp. 338–344. Association for Computational Linguistics, Beijing, China (Jul 2015), https://www.aclweb.org/anthology/K15-1037

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

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)

Vulić, I., Moens, M.F.: Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 363–372. SIGIR '15, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2766462.2767752

Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. 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. 1112–1122. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://www.aclweb.org/anthology/N18-1101

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing (2019)

Zhong, Z., Ng, H.T.: It makes sense: A wide-coverage word sense disambiguation system for free text. In: Proceedings of the ACL 2010 System Demonstrations. pp. 78–83. Association for Computational Linguistics, Uppsala, Sweden (Jul 2010), https://www.aclweb.org/anthology/P10-4014