Transfer Learning and Augmentation for Word Sense Disambiguation

Harsh Kohli(✉)

SalesKen, Bengaluru, India
harshkohli@salesken.ai

Abstract. Many downstream NLP tasks have shown significant improvement through continual pre-training, transfer learning and multi-task learning. State-of-the-art approaches in Word Sense Disambiguation today benefit from some of these approaches in conjunction with information sources such as semantic relationships and gloss definitions contained within WordNet. Our work builds upon these systems and uses data augmentation along with extensive pre-training on various different NLP tasks and datasets. Our transfer learning and augmentation pipeline achieves state-of-the-art single model performance in WSD and is at par with the best ensemble results.

Keywords: Word Sense Disambiguation · Multi-task training · Transfer learning

1 Introduction

Word Sense Disambiguation or WSD is the task of gleaning the correct sense of an ambiguous word given the context in which it was used. It is a well-studied problem in NLP and has seen several diversified approaches over the years including techniques leveraging Knowledge-Based Systems, Supervised learning approaches and, more recently, end-to-end deep learnt models. WSD has found application in various kinds of NLP systems such as Question Answering, IR, and Machine Translation.

WordNet 3.0 is the most popular and widely used sense inventory that consists of over 109k synonym sets or synsets and relationships between them such as hypernym, anotnym, hyponym, entailment etc. Most training and evaluation corpora used in supervised systems today consist of sentences where words are manually annotated and mapped to a particular synset in WordNet. We use these sources in addition to other publicly available datasets to tune our model for this task. Through transfer learning from these datasets and other augmentation and pre-processing techniques we achieve state-of-the-art results on standard benchmarks.

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, pp. 303–311, 2021.
https://doi.org/10.1007/978-3-030-72240-1_29

2 Related Work

Traditional approaches to WSD relied primarily on Knowledge-Based Systems. Lexical similarity over dictionary definitions or Gloss for each synset was first used in [10] to estimate the correct sense. Graph based approaches such as [18] were also proposed which leverage structural properties of lexico-semantic sources treating the knowledge graph as a semantic network. One major advantage of using such unsupervised techniques was that they eliminated the need of having large annotated training corpora. Since annotation is expensive given the large number of fine-grained word senses, such methods were the de facto choice for WSD systems. Recently, however, approaches for semi-automatic [27] and automatic [21] sense annotation have been proposed to partially circumvent the problem of manually annotating a sizeable training set.

Supervised methods, on the other hand, relied on a variety of hand-crafted features such as a neighbouring window of words and their corresponding part of speech (POS) tags etc. Commonly referred to as word expert systems, they involved training a dedicated classifier for each individual lemma [34]. The default or first sense was usually returned when the target lemma was not seen during training. While these were less practical in real application, they often yielded better results on common evaluation sets.

[8] and [24] were the first neural architectures for WSD which consisted of Bidirectional LSTM models and Seq2Seq Encoder-Decoder architectures with attention. These architectures optionally included lexical and POS features which yielded better results. Due to strong performance of contextual embeddings such as BERT [3] on various NLP tasks, recent approaches such as [30] and [5] have used these to achieve significant gains in WSD benchmarks. We leverage the ideas presented in GlossBERT [5] and improve upon the results with a multi-task pre-training procedure and greater semantic variations in the train dataset through augmentation techniques.

3 Data Preparation Pipeline

3.1 Source Datasets

We use the largest manually annotated WSD corpus SemCor 3.0 [17] consisting of over 226k sense tags for training our models. In keeping with most neural architectures today such as [14], we use the SemEval-2007 corpus [22] as our dev set and SemEval-2013 [20], SemEval-2015 [19], Senseval-2 [4], and Senseval-3 [26] as our test sets.

3.2 Data Preprocessing

GlossBERT [5] utilizes context gloss pairs with weak supervision to achieve state-of-the-art single model performance on the evaluation sets. We follow the same pre-processing procedure as GlossBERT. The context sentence along with each

of the gloss definitions of senses of the target word are considered as a pair. Thus, for a sentence containing an ambiguous word with N senses, we consider all N senses with as many sentence pairs. Only the correct sense is marked as a positive sample while all others are considered negative inputs to our pairwise sentence classifier. As this formulation relies on the gloss definition of a synset and not just the synset tag or key, it is more robust to keys that do not occur or are under-represented in training.

Context Sentence	Gloss Definition	Label
How long has it been since you reviewed the " objectives " of your benefit and service program ?	objectives : the lens or system of lenses in a telescope or microscope that is nearest the object being viewed	0
How long has it been since you reviewed the " objectives " of your benefit and service program ?	objectives : the goal intended to be attained (and which is believed to be attainable)	1
How long has it been since you reviewed the " objectives " of your benefit and service program ?	objectives : undistorted by emotion or personal bias; based on observable phenomena	0
How long has it been since you reviewed the " objectives " of your benefit and service program ?	objectives : serving as or indicating the object of a verb or of certain prepositions and used for certain other purposes	0
How long has it been since you reviewed the " objectives " of your benefit and service program ?	objectives : belonging to immediate experience of actual things or events	0
How long has it been since you reviewed the " objectives " of your benefit and service program ?	objectives : emphasizing or expressing things as perceived without distortion of personal feelings, insertion of fictional matter, or interpretation	0

Fig. 1. Context-Gloss Pairs with Weak Supervision

Figure 1 above shows an example of context-gloss pairs for a single context sentence with the target word - objectives. The highlighted text represent the weak supervised signals which help identify the target word both in the gloss definition, as well as in the context sentence. In the context sentence, the target word may appear more than once, and the signal helps associate each occurrence with the definition independently.

3.3 Data Augmentation

Given the large number of candidate synsets for each target lemma, the train dataset has a large class imbalance. The ratio of negative samples to positives is nearly 8:1. Rather than adopting a simple oversampling strategy, we use data augmentation through back translation. Back translation is a popular method for generating paraphrases involving translating a source sentence to one of several target languages and then translating the sentence back into the source language. Approaches described in [16,23,32] have successfully leveraged modern Neural Machine Translation systems to generate paraphrases for a variety of tasks. We use this technique to introduce greater diversity and semantic variation in our training set and augment examples in our minority class.

The Transformers library [33] provides MarianMT models [7] for translation to and from several different languages. Each model is a 6-layer transformer [29] encoder-decoder architecture. For best results, we select from a number of high-resource languages such as French, German etc. and apply simple as well as chained back-translation (e.g. English - Spanish - English - French - English). From our pool of back-translated sentences, we retain sentences where the target word occurs exactly once in the original as well as back-translated sentence. This

way, we generate several paraphrased examples for each positive example in our train set. We randomly select n augmented samples for each original sample at train time, where n was treated as a hyper-parameter during our training experiments. We achieve best results when $n = 3$.

4 Model

We use the MT-DNN [12] architecture for training our model. The network consists of shared layers and task-specific layers. Through cross-task training, the authors demonstrate how the shared layers of the network learn more generalized representations and are better suited to adapt to new tasks and domains. Multi-task learning using large amount of labelled data across tasks has a regularization effect on the network and the model is able to better generalize to new domains with relatively fewer labelled training examples than simple pre-trained BERT. It is this property of MT-DNN that we leverage to improve performance on WSD.

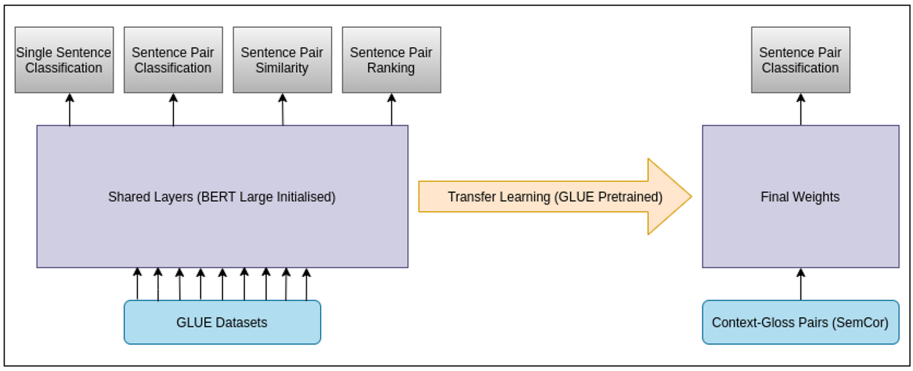

Fig. 2. Pre-training and Tuning methodology

The pre-training procedure for MT-DNN is similar to that of BERT which used two supervised tasks - masked LM and next sentence prediction. Using BERT Large model (24 layers, 1024 dim, 335 m trainable parameters) as our base model, we then tune on all tasks in the GLUE benchmark [31]. While [5] reported better performance using BERT base (12 layers, 768 dim, 110 m trainable parameters), we found that the larger BERT model performed significantly better in our experiments. We attribute this behaviour to our pre-training procedure which learns better, more generalized representations thus preventing a larger, more expressive model from overfitting on the train dataset.

Four different task-specific output layers are constructed corresponding to single sentence classification, pairwise text similarity, pairwise text classification, and pairwise text ranking. These are illustrated in Fig. 2. Learning objectives

differ for each task - single-sentence and pairwise classification tasks are optimized using cross-entropy loss, pairwise text similarity is optimized on the mean squared error between the target similarity value and semantic representations of each of the sentences in the input pair, and pairwise text ranking follows the pairwise learning-to-rank paradigm in minimizing the negative log likelihood of a positive example given a list of candidates [2]. The pairwise text classification output layer uses a stochastic answer network (SAN) [11] which maintains a memory state and employs K-step reasoning to iteratively improve upon predictions. We use the same pairwise classification head when tuning the network for our WSD task. At inference time, we run context-gloss pairs for each sense of the target lemma and the candidate synset with the highest score is considered the predicted sense.

5 Implementation Details

Examples from each of the 9 datasets in GLUE are input to the network and passed to the correct output layer given the task-type. 5 epochs of pre-training are thus carried out using GLUE data. The best saved checkpoint is then selected and, thereafter, context-gloss pairs as described above are input to the model for tuning on WSD. Model weights of shared layers are carried over from multi-task training on GLUE. Adamax [9] optimizer is used to tune the weights and a low learning rate of 2e−5 is used to facilitate a slower, but smoother convergence. A batch size of 256 is maintained and the architecture is tuned on 8x Tesla V100 GPU's with 16 GB of VRAM each for a total of 128 GB GPU memory.

6 Results

We summarize the results of our experiments in Table 1. We compare our results against the Most Frequent Sense Baseline as well as different approaches, Knowledge Based - Lesk (ext+emb) [1] and Babelfly [18], Word-Expert Supervised Systems - IMS [34] and IMS+emb[6], Neural Models - Bi-LSTM [8], Bi-LSTM + att + lex +pos [24], CAN/HCAN [14], GAS [15], SemCar/SemCor+WNGC, hypernyms [30] and GlossBERT [5]. We exclude results from ensemble systems marked in Table 1 as these results were obtained using a geometric mean of predictions across 8 independent models. We achieve the best results for any single model across all evaluation sets and POS types.

While [30] supplement their train corpus with the Wordnet Gloss Corpus (WNGC) and also use 8 different models for their ensemble, our overall results are at par with theirs on test datasets and slightly better on the dev set. The fact that such results were achieved with fewer training examples (without the use of WNGC) further enforces the generalization and domain adaptation capabilities of our pre-training methodology.

Table 1. Final Results. * Result excluded from consideration as it uses an ensemble

System	SE07	SE2	SE3	SE13	SE15	Noun	Verb	Adj	Adv	All
MFS Baseline	54.5	65.6	66.0	63.8	67.1	67.7	49.8	73.1	80.5	65.5
$\text{Lesk}_{\text{ext+emb}}$	56.7	63.0	63.7	66.2	64.6	70.0	51.1	51.7	80.6	64.2
Babelfly	51.6	67.0	63.5	66.4	70.3	68.9	50.7	73.2	79.8	66.4
IMS	61.3	70.9	69.3	65.3	69.5	70.5	55.8	75.6	82.9	68.9
$\text{IMS}_{\text{+emb}}$	62.6	72.2	70.4	65.9	71.5	71.9	56.6	75.9	84.7	70.1
Bi-LSTM	–	71.1	68.4	64.8	68.3	69.5	55.9	76.2	82.4	68.4
$\text{Bi-LSTM}_{\text{+att.+LEX+POS}}$	64.8	72.0	69.1	66.9	71.5	71.5	57.5	75.0	83.8	69.9
GAS_{ext}(Linear)	–	72.4	70.1	67.1	72.1	71.9	58.1	76.4	84.7	70.4
GAS_{ext}(Concatenation)	–	72.2	70.5	67.2	72.6	72.2	57.7	76.6	85.0	70.6
CAN	–	72.2	70.2	69.1	72.2	73.5	56.5	76.6	80.3	70.9
HCAN	–	72.8	70.3	68.5	72.8	72.7	58.2	77.4	84.1	71.1
SemCor,hyp	–	–	–	–	–	-	–	–	–	75.6
SemCor,hyp(ens)*	69.5	77.5	77.4	76.0	78.3	79.6	65.9	79.5	85.5	76.7
SemCor+WNGC,hyp	–	–	–	–	–	–	–	–	–	77.1
SemCor+WNGC,hyp(ens)*	73.4	79.7	77.8	78.7	82.6	81.4	68.7	83.7	85.5	79.0
BERT(Token-CLS)	61.1	69.7	69.4	65.8	69.5	70.5	57.1	71.6	83.5	68.6
GlossBERT(Sent-CLS)	69.2	76.5	73.4	75.1	79.5	78.3	64.8	77.6	83.8	75.8
GlossBERT(Token-CLS)	71.9	77.0	75.4	74.6	79.3	78.3	66.5	78.6	84.4	76.3
GlossBERT(Sent-CLS-WS)	72.5	77.7	75.2	76.1	80.4	79.3	66.9	78.2	86.4	77.0
MTDNN+Gloss	**73.9**	**79.5**	**76.6**	**79.7**	**80.9**	**81.8**	**67.7**	**79.8**	**86.5**	**79.0**

7 Conclusion and Future Work

We use the pre-processing steps and weak-supervision over context-gloss pairs as described in [5] and improve upon the results through simple and chained back-translation as a means of data augmentation and multi-task training and transfer learning from different data sources. Better and more generalized representations achieved by leveraging the GLUE datasets allows us to train a larger model with nearly thrice as many trainable parameters. Through these techniques we are able improve upon existing SOTA on standard benchmark.

Additional data from WNGC or OMSTI [27] has shown to aid model performance in various systems and could be incorporated in training. Recent work such as [28] indicates that cost-sensitive training is often effective when training BERT when there is a class imbalance. Given the nature of the problem, a triplet loss function similar to [25] could be used to further improve performance. Online hard or semi-hard sampling strategies could be experimented with to sample the negative sysnets. Finally, RoBERTa [13] has shown improved performance on many NLP tasks and could be used as a base model that is input to our multi-task pre-training pipeline. All of these techniques could be used in conjunction with our context-gloss pairwise formulation to improve performance further.

References

1. Basile, P., Caputo, A., Semeraro, G.: An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, pp. 1591–1600. Dublin City University and Association for Computational Linguistics, August 2014. https://www.aclweb.org/anthology/C14-1151
2. Burges, C., et al.: Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine Learning, ICML 2005, pp. 89–96, New York, NY, USA. Association for Computing Machinery (2005). https://doi.org/10.1145/1102351.1102363, https://doi.org/10.1145/1102351.1102363
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:abs/1810.04805 (2019)
4. Edmonds, P., Cotton, S.: SENSEVAL-2: overview. In: Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France, pp. 1–5. Association for Computational Linguistics, July 2001. https://www.aclweb.org/anthology/S01-1001
5. Huang, L., Sun, C., Qiu, X., Huang, X.: Glossbert: Bert for word sense disambiguation with gloss knowledge. arXiv:abs/1908.07245 (2019)
6. Iacobacci, I., Pilehvar, M.T., Navigli, R.: 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), Berlin, Germany, pp. 897–907. Association for Computational Linguistics, August 2016. https://doi.org/10.18653/v1/P16-1085, https://www.aclweb.org/anthology/P16-1085
7. Junczys-Dowmunt, M., et al.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations, Melbourne, Australia, pp. 116–121. Association for Computational Linguistics, July 2018. http://www.aclweb.org/anthology/P18-4020
8. Kågebäck, M., Salomonsson, H.: Word sense disambiguation using a bidirectional LSTM. In: Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V), Osaka, Japan, pp. 51–56. The COLING 2016 Organizing Committee, December 2016. https://www.aclweb.org/anthology/W16-5307
9. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations, December 2014
10. Lesk, M.E.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: SIGDOC 1986 (1986)
11. Liu, X., Duh, K., Gao, J.: Stochastic answer networks for natural language inference, April 2018
12. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 4487–4496. Association for Computational Linguistics, July 2019. https://www.aclweb.org/anthology/P19-1441
13. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019). http://arxiv.org/abs/1907.11692

14. Luo, F., Liu, T., He, Z., Xia, Q., Sui, Z., Chang, B.: 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, Brussels, Belgium, pp. 1402–1411. Association for Computational Linguistics, October–November 2018. https://doi.org/10.18653/v1/D18-1170, https://www.aclweb.org/anthology/D18-1170
15. Luo, F., Liu, T., Xia, Q., Chang, B., Sui, Z.: Incorporating glosses into neural word sense disambiguation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia , pp. 2473–2482. Association for Computational Linguistics, July 2018. https://doi.org/10.18653/v1/P18-1230, https://www.aclweb.org/anthology/P18-1230
16. Mallinson, J., Sennrich, R., Lapata, M.: Paraphrasing revisited with neural machine translation. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, pp. 881–893. Association for Computational Linguistics, April 2017. https://www.aclweb.org/anthology/E17-1083
17. Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T.: A semantic concordance. In: Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21–24, 1993 (1993). https://www.aclweb.org/anthology/H93-1061
18. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Trans. Assoc. Comput. Linguist. **2**, 231–244 (2014)
19. 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), Denver, Colorado, pp. 288–297. Association for Computational Linguistics, June 2015. https://doi.org/10.18653/v1/S15-2049, https://www.aclweb.org/anthology/S15-2049
20. 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), Atlanta, Georgia, USA, pp. 222–231. Association for Computational Linguistics, June 2013. https://www.aclweb.org/anthology/S13-2040
21. Pasini, T., Navigli, R.: Train-o-matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data, pp. 78–88, January 2017. https://doi.org/10.18653/v1/D17-1008
22. Pradhan, S., Loper, E., Dligach, D., Palmer, M.: SemEval-2007 task-17: English lexical sample, SRL and all words. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, pp. 87–92. Association for Computational Linguistics, June 2007. https://www.aclweb.org/anthology/S07-1016
23. Prakash, A., Hasan, S.A., Lee, K., Datla, V., Qadir, A., Liu, J., Farri, O.: Neural paraphrase generation with stacked residual LSTM networks. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, pp. 2923–2934. The COLING 2016 Organizing Committee, December 2016. https://www.aclweb.org/anthology/C16-1275
24. Raganato, A., Delli Bovi, C., Navigli, R.: Neural sequence learning models for word sense disambiguation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1156–1167, Copenhagen, Denmark. Association for Computational Linguistics, September 2017. https://doi.org/10.18653/v1/D17-1120, https://www.aclweb.org/anthology/D17-1120

25. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)
26. Snyder, B., Palmer, M.: The English all-words task. In: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, pp. 41–43. Association for Computational Linguistics, July 2004. https://www.aclweb.org/anthology/W04-0811
27. 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, Beijing, China, pp. 338–344. Association for Computational Linguistics, July 2015. https://doi.org/10.18653/v1/K15-1037, https://www.aclweb.org/anthology/K15-1037
28. Tayyar Madabushi, H., Kochkina, E., Castelle, M.: Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, Hong Kong, China, pp. 125–134. Association for Computational Linguistics, November 2019. https://doi.org/10.18653/v1/D19-5018, https://www.aclweb.org/anthology/D19-5018
29. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008. Curran Associates, Inc. (2017). http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
30. Vial, L., Lecouteux, B., Schwab, D.: Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. CoRR abs/1905.05677 (2019). http://arxiv.org/abs/1905.05677
31. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, pp. 353–355. Association for Computational Linguistics, November 2018. https://doi.org/10.18653/v1/W18-5446, https://www.aclweb.org/anthology/W18-5446
32. Wieting, J., Gimpel, K.: ParaNMT-50M: pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, pp. 451–462. Association for Computational Linguistics, July 2018. https://doi.org/10.18653/v1/P18-1042, https://www.aclweb.org/anthology/P18-1042
33. Wolf, T., et al.: Huggingface's transformers: state-of-the-art natural language processing, arXiv:abs/1910.03771 (2019)
34. 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, Uppsala, Sweden, pp. 78–83. Association for Computational Linguistics, July 2010. https://www.aclweb.org/anthology/P10-4014