# Integrating Personalized PageRank into Neural Word Sense Disambiguation

**Ahmed ElSheikh** **Michele Bevilacqua** **Roberto Navigli**
Sapienza NLP Group
Sapienza University of Rome
firstname.lastname@uniroma1.it

## Abstract

Neural Word Sense Disambiguation (WSD) has recently been shown to benefit from the incorporation of pre-existing knowledge, such as that coming from the WordNet graph. However, state-of-the-art approaches have been successful in exploiting only the *local* structure of the graph, with only close neighbors of a given synset influencing the prediction. In this work, we improve a classification model by recomputing logits as a function of both the vanilla independently produced logits and the *global* WordNet graph. We achieve this by incorporating an online neural approximated PageRank, which enables us to refine edge weights as well. This method exploits the *global* graph structure while keeping space requirements linear in the number of edges. We obtain strong improvements, matching the current state of the art. Code is available at https://github.com/SapienzaNLP/neural-pagerank-wsd.

## 1 Introduction

Word Sense Disambiguation (WSD) is a task with a long history in Natural Language Processing (Bevilacqua et al., 2021). It addresses the pervasive phenomenon of (lexical) ambiguity, whereby the same polysemous word or expression conveys different meanings in different contexts. Natural Language Processing (NLP) tends to maintain the working assumption that word meaning can be discretized in a finite number of classes, thus casting polysemy resolution as a multi-class classification problem, where the classes, i.e., the senses, are specific to a word. Senses are registered in a dictionary-like resource called the sense inventory. In English WSD the sense inventory is virtually always WordNet (Miller et al., 1990), which, for example, lists separately the *fish* and *musical instrument* senses of the word "bass".

The WSD task is currently dominated by supervised methods (Yap et al., 2020; Blevins and Zettlemoyer, 2020). These methods, thanks, inter alia, to the game-changing effect of pretrained language models, have widened the margin over so-called knowledge-based approaches (Agirre and Soroa, 2009; Moro et al., 2014; Scozzafava et al., 2020), which usually disambiguate using only global graph information – a source of information that, however, is not easy to integrate explictly into supervised WSD. Although there have been a few successful approaches integrating graphs into a standard neural classification architecture (Conia and Navigli, 2021; Bevilacqua and Navigli, 2020), such methods only exploit the *local* relational structure, leaving the *global* structure unused. Thus, while the model is able to directly utilize the explicit knowledge that a *cairn* is a *terrier*, it is not able to also utilize the fact that, following the hyponymy relation, a *cairn* is also a *dog*.

In this paper we propose a method for integrating global graph information into neural supervised WSD through a Personalized PageRank (Page et al., 1999, PPR) approximation, blurring the distinction between knowledge-based and supervised methods. We achieve this by generalizing the logit aggregation scheme used on top of a feedforward classifier by Bevilacqua and Navigli (2020). Our proposed method is simple and extensible, and could be adapted to work with other classifiers as well. We match the performance of the current state of the art in WSD (Blevins and Zettlemoyer, 2020) while using only 35% (60M) of the trainable parameter count. When training on additional data, our technique outperforms the previous state of the art in that setting (Conia and Navigli, 2021).

## 2 Method

In this section we explain how our method, building on top of previous approaches to WSD.



*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9092–9098
November 7–11, 2021. ©2021 Association for Computational Linguistics

### 2.1 Neural WSD

The most popular formulation for WSD employs token-level classifiers, which encode each token $t_i$ into a vector $e_i$ using a sequence classifier, e.g. an LSTM (Raganato et al., 2017b) or, more recently, pre-trained Transformer-based contextualized embeddings (Hadiwinoto et al., 2019). The vector is then fed to one or more feedforward (FFN) layers, producing a softmax-normalized probability distribution over all the output classes, i.e., all the synsets (groups of senses sharing the same meaning) in WordNet:

$$\begin{aligned} P(s|t_i) &= \frac{\exp(\text{FFN}(e_i))_s}{\sum_{s' \in \mathcal{I}(\cdot)} \exp(\text{FFN}(e_i))_{s'}} \\ &= \text{softmax}(\text{FFN}(e_i))_s \end{aligned} \tag{1}$$

where $\mathcal{I}$ is an inventory function that returns the set of possible synsets for a token, and $\mathcal{I}(\cdot)$ denotes the set of all WordNet synsets. At test time, the predicted synset $\hat{s}_i$ is the one with the highest probability, searching only among the set of synsets that are consistent with the current token ($\mathcal{I}(t_i)$):

$$\hat{s}_i = \underset{s' \in \mathcal{I}(t_i)}{\text{argmax}}\, P(s'|t_i) \tag{2}$$

This formulation is very straightforward, but also wasteful, as scores are computed for "impossible" synsets, i.e. those $\notin \mathcal{I}(t_i)$.

### 2.2 EWISER

Recently, there has been a trend towards the inclusion of knowledge from the WordNet sense inventory in WSD. While many successful approaches have exploited glosses (Huang et al., 2019; Blevins and Zettlemoyer, 2020), the use of relational information, i.e. the graph structure of WordNet, has mostly been exploited by so-called knowledge-based algorithms, i.e., those that make use of no corpus supervision at all. However, one recent approach (Bevilacqua and Navigli, 2020, EWISER) has made the use of graph relations part of its core method: it employs trainable edge weights to aggregate scores of related synsets together, thus taking advantage of the scores over the whole vocabulary:

$$\begin{aligned} Z^{(0)}_{t_i,s} &= \text{FFN}(e_i) \\ Z^{(1)}_{t_i,s} &= Z^{(0)}_{t_i,s} + \sum_{s' \in E_{in}(s)} w(s', s) \\ P(s|t_i) &= \text{softmax}(Z^{(1)}_{t_i})_s \end{aligned} \tag{3}$$

where $E_{in}(s)$ is the set of all the synsets $s' \in V$ (i.e., the set of nodes), such that there is an edge from $s'$ to $s$, and $w(s', s)$ is the weight of the edge. Bevilacqua and Navigli (2020) batch the computation by encoding all edges in a (sparse) adjacency matrix $A \in \mathbb{R}^{|\mathcal{I}(\cdot)| \times |\mathcal{I}(\cdot)|}$, where $A_{s',s} = 0$ unless $s' \in E_{in}(s)$, in which case $A_{s',s} = \frac{1}{|E_{in}(s)|}$:

$$Z^{(1)} = Z^{(0)} + A^T Z^{(0)} \tag{4}$$

In training, only non-zero $A$ weights are updated.

EWISER provides an elegant way to incorporate graph knowledge within a token tagger architecture. However, it also has some evident limitations. First, the model is only able to incorporate local neighbourhood, as only paths of length 1 are considered. While one could incorporate paths of arbitrary length by augmenting $E_{in}(s)$ with all nodes connected to $s$ by at most $k$ steps, this solution is in practice not very scalable, as it would rapidly densify $A$, increasing the number of parameters way beyond what is reasonable: if we were to have a fully connected graph for WordNet synsets, it would have more than 6.9 billion parameters.

### 2.3 Integrating PageRank

We propose improving the logit aggregation mechanism of EWISER (Eq. 4) by applying the $A$ matrix $K$ times. Each new iteration makes it such that progressively more distant neighbours affect the classification score. However, given a sufficiently large value of $K$, this would have the effect that the contribution of the original scores becomes increasingly smoothed out.

To solve this issue we exploit the connection between the logit aggregation step and the PPR algorithm (Page et al., 1999), which uses both edge information and a personalization vector ($z$, a prior on node importance before taking into account edges) to produce a distribution over nodes. One of the ways to "solve" the PageRank uses the so-called power iteration method, whereby $z$ is repeatedly multiplied by $A$:

$$\begin{aligned} PR^{(0)} &= z \\ PR^{(n)} &= \alpha z + (1 - \alpha) A^T PR^{(n-1)} \end{aligned} \tag{5}$$

where $\alpha$ is the so-called teleport probability, used to interpolate between $z$ and the current iteration scores – saving some probability mass from the former. In our case for each instance to classify

the personalization vector is set to be equal to logit scores, i.e. the raw scores before applying softmax:

$$\begin{aligned} Z_{t_i}^{(0)} &= \text{FFN}(e_i) \\ Z^{(k)} &= \alpha Z^{(0)} + (1-\alpha) A^T Z^{(k-1)} \\ P(s|t_i) &= \text{softmax}(Z_{t_i}^{(K)})_s \end{aligned} \quad (6)$$

Differently from vanilla PPR, we do not check for convergence, but treat $K$ as a hyperparameter; also, we do not normalize the personalization vector into a probability distribution, as preliminary experiments showed that normalization does not affect the final classification scores significantly. Our approach is related to that of Klicpera et al. (2019), who use a neural approximation based on topic-sensitive PageRank (Haveliwala, 2002) instead of PPR, and apply it to the task of node classification in citation graphs. Differently from them, however, we exploit the graph structure of the output, not that of the input.

## 3 Experiments

**Setting** We evaluate our proposed addition to the WSD task by employing it on top of the simple feedforward classifier baseline (taking as input frozen BERT large embeddings) used by Conia and Navigli (2021). We train the model with vanilla categorical cross-entropy. Following Bevilacqua and Navigli (2020) we use SensEmBERT (Scarlini et al., 2020) and LMMS (Loureiro and Jorge, 2019) embeddings to initialize the output embeddings (i.e., the last transformation matrix of the FFN) for, respectively, nominal and all other synsets. The evaluation measure is the $F_1$ score. We compare against EWISER, but also report results for other recent high-performing methods from the literature, both sequence-tagging (Huang et al., 2019; Yap et al., 2020) and token-tagging (Blevins and Zettlemoyer, 2020; Conia and Navigli, 2021).[1]

**Data** As usual in the WSD literature, we train our model on SemCor (Miller et al., 1994), i.e. the largest available manually semantically annotated corpus; following Bevilacqua and Navigli (2020) we also add synthetic instances built by prepending lemmas to the corresponding WordNet synset definition – thus injecting gloss information that is used by other state-of-the-art models (Huang et al., 2019; Yap et al., 2020; Blevins and Zettlemoyer, 2020). In order to test the performance upper bounds we also experiment separately with adding to the training set the so-called WordNet Tagged Glosses corpus[2] (WNTG), which contains a large number of additional gold and silver annotations.

We evaluate on the framework for English all-words WSD made available by Raganato et al. (2017a),[3] which includes Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013), and SemEval-2015 (Moro and Navigli, 2015). We use SemEval-2007 as our development set, and report results on the concatenation of all other datasets in the framework (ALL$^-$); following common practice, we also report results on SemEval-2007 + ALL$^-$ (ALL), even though the former is the development set.

| $K$ | **Connected paths** (%) |
|---|---|
| 4 | 11.083 |
| 5 | 33.538 |
| 6 | 60.543 |
| 7 | 81.529 |
| 8 | 92.815 |
| 9 | 97.579 |
| 10 | 99.313 |
| 11 | 99.849 |
| 12 | 99.967 |
| 13 | 99.997 |

Table 1: Percentage of paths between any two connected synsets in WordNet whose length is less than or equal to $K$, for different values of $K$.

**Graph** We train models with various values of the power iteration parameter $K$: (i) $K = 0$, which corresponds to the baseline with no logit aggregation scheme; (ii) $K = 1$, which is similar to EWISER, but using $\alpha$-weighted interpolation between original logits and aggregated ones; (iii) $K = 10$, which corresponds to a range greater than or equal to the length of around 99% of paths between any two connected nodes in the graph. We report the path length statistics in Table 1.

We build $A$ by including different sets of edges

[1] We have not included in the comparison work appearing at the time of or later than our submission (Barba et al., 2021a; Wang and Wang, 2021; Barba et al., 2021b)

[2] https://wordnetcode.princeton.edu/glosstag.shtml

[3] http://lcl.uniroma1.it/wsdeval/

| | Model | ALL$^-$ | ALL | S2 | S3 | S7 | S13 | S15 | N | V | A | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *SemCor only* | Huang et al. (2019) | 77.0 | 76.8* | 77.7 | 75.2 | 72.5 | 76.1 | 80.4 | 79.8 | 67.1 | 79.6 | 87.4 |
| | Conia and Navigli (2021) | 77.8* | 77.6 | 78.4 | 77.8 | 72.2 | 76.7 | 78.2 | 80.1 | 67.0 | 80.5 | 86.2 |
| | Bevilacqua and Navigli (2020) | 78.8* | 78.3 | 78.9 | 78.4 | 71.0 | 78.9 | 79.3 | 81.7 | 66.3 | 81.2 | 85.8 |
| | Yap et al. (2020) | 79.1* | 78.7 | 79.9 | 77.4 | 73.0 | 78.2 | 81.8 | 81.2 | 68.8 | 81.5 | 88.2 |
| | Blevins and Zettlemoyer (2020) | 79.3* | **79.0** | 79.4 | 77.4 | 74.5 | 79.7 | 81.7 | 81.4 | 68.5 | 83.0 | 87.9 |
| | Ours ($K = 0$) | 77.1 | 76.5 | 77.6 | 75.4 | 68.8 | 76.9 | 79.3 | 79.5 | 66.0 | 78.1 | 85.0 |
| | Ours ($K = 1$) | 78.5$\star$ | 78.1 | 78.5 | 78.1 | 71.4 | 77.7 | 80.8 | 80.6 | 68.6 | 80.5 | 86.7 |
| | Ours ($K = 10$) | **79.4** | 78.9 | 78.7 | 78.9 | 71.6 | 80.0 | 81.2 | 82.3 | 67.8 | 80.7 | 86.1 |
| | Ours ($K = 10$) + SyntagNet | 79.2 | 78.7 | 78.9 | 77.8 | 72.1 | 79.0 | 82.4 | 81.6 | 68.5 | 80.7 | 86.4 |
| | Ours (PPR top-10, $K = 1$) | 78.2 | 77.6 | 78.0 | 77.6 | 70.1 | 77.9 | 79.9 | 80.7 | 67.0 | 79.5 | 85.5 |
| *+WNTG* | Bevilacqua and Navigli (2020) | 80.4* | 80.1 | 80.8 | 79.0 | 75.2 | 80.7 | 81.8 | 82.9 | 69.4 | 83.7 | 85.5 |
| | Conia and Navigli (2021) | 80.5* | 80.2 | 80.4 | 77.8 | 76.2 | 81.8 | 83.3 | 82.9 | 70.3 | 83.4 | 85.5 |
| | Ours ($K = 0$) | 78.6 | 78.1 | 78.4 | 78.3 | 70.9 | 78.1 | 80.1 | 80.9 | 67.8 | 80.5 | 86.7 |
| | Ours ($K = 1$) | 80.2$\star$ | 79.8 | 79.5 | 77.7 | 74.2 | 81.3 | 84.5 | 82.6 | 70.1 | 82.2 | 85.5 |
| | Ours ($K = 10$) | 80.7 | 80.3 | 80.3 | 79.1 | 74.3 | 81.3 | 83.4 | 83.2 | 69.1 | 84.7 | 85.3 |
| | Ours ($K = 10$) + SyntagNet | **81.0** | **80.6** | 80.5 | 78.5 | 74.3 | 81.9 | 85.3 | 83.7 | 69.9 | 83.1 | 86.4 |

Table 2: We report results ($F_1$) of the WSD experiments. Results on concatenated datasets (ALL$^-$/ALL); Senseval-2/3 (S2/S3); SemEval-2007/2013/2015; part-of-speech breakdown on ALL: nouns (N), verbs (V), adjectives (A), adverbs (R). Models grouped in the same row block are mutually comparable: 1) models trained on just SemCor, 2) models trained on SemCor and WNTG. *: computed from the reported results. $\star$ : highest $F_1$ that is statistically different from the best one (McNemar's test with $p = 0.05$).

from WordNet, i.e., *hyponymy*, *hypernyms*, *similarity*, *derivationally related*, and *verb group*. The weight $A_{s',s}$ from synset $s'$ to $s$ is initialized as $1/|E_{in}(s)|$ where $E_{in}(s)$ is the set of all $s'$ s.t. the edge $\langle s', s\rangle$ is in the graph. Additionally, we experiment with including edges from SyntagNet (Maru et al., 2019), a resource that includes edges representing semantic collocations.[4] Collocational information is orthogonal to that contained in WordNet, providing paths between regions of the WordNet graph that would otherwise be distant or disconnected.

Finally, to check whether it is feasible to include global information by *precomputing* the PPR instead of approximating it in the network forward pass, we experiment with a baseline approach where we directly initialize $A_s^T$ with a PPR distribution, built using WordNet as the starting graph and a one-hot vector $z$ (with $z_s = 1$) as personalization. To keep $A$ manageable, we cut the distribution in $A_s^T$ the top 10 ranks and renormalize to sum to 1. We then train a $K = 1$ model on this graph.

**Hyperparameters** Models are trained for 30 epochs, using early stopping with patience set to 5, feeding data in batches of 128 sentences. The optimizer used is Adam (Kingma and Ba, 2015) with a learning rate of $10^{-4}$ which decays linearly to $10^{-7}$. We set the teleport probability $\alpha$ to 0.15, a common default value. We do not tune $K$, nor other hyperparameters, which we take from the configuration of Bevilacqua and Navigli (2020). The development set is only used for early stopping and to select the best epoch of the run.

## 4 Results

We report results of our experiments in Table 2. When training on SemCor, the performances of our model increase steadily from $K = 0$ to $K = 1$ (+1.4 on ALL$^-$), and from $K = 1$ to $K = 10$ (+0.9 on ALL$^-$), while SyntagNet edges do not seem to boost the results over the use of the simple WordNet graph. The precomputed PPR graph baseline obtains much lower results than $K = 10$. The reason is probably that the baseline needs to keep a fixed number of incoming edges, missing potentially useful information on a structurally longer range, while our $K = 10$ model has no such requirement. On the overall evaluation (ALL$^-$) the results of the $K = 10$ model also match (and even slightly outperform) those of the current state of the art, i.e., BEM (Blevins and Zettlemoyer, 2020), all while using around 35% of the trainable parameters. In fact, apart from the parameters in $A$, our model is a simple feedforward network on top of BERT, while BEM uses two jointly fine-tuned encoders, one to encode the context and the other to encode the definitions.

When adding WNTG to the training corpus, the same trend of increasing performance along with increasing $K$ appears. Moreover, our best model

[4] http://syntagnet.org/

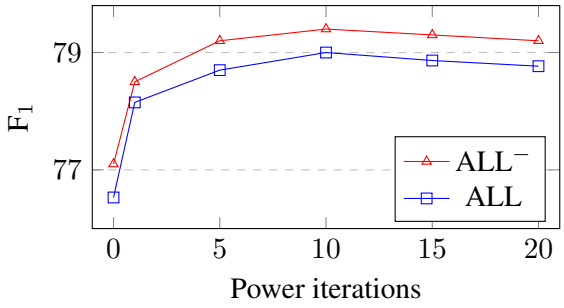


Figure 1: Performances on ALL⁻/ALL with $K \in \{0, 1, 5, 10, 15, 20\}$ of our SemCor model (w/o SyntagNet).

| Model | Seen | Unseen |
|---|---|---|
| Ours ($K = 1$) | 93.9 | 65.1 |
| Ours ($K = 10$) | **94.1** | **67.0** |

Table 3: Results of model trained on SemCor (for $K = 1$ and $K = 10$), on the subset of ALL⁻ such that the MFS is among gold synsets (**MFS**) and on its complement in ALL⁻ (**Unseen**). $F_1$ is reported.

outperforms the previous state of the art (Conia and Navigli, 2021) by 0.5 points on ALL⁻. In this new setting using SyntagNet edges is beneficial, probably because now that many more senses have meaningful occurrences in the larger training set, there is less noise in the tail of the $Z^{(0)}$ (unnormalized) distribution, and more synsets can affect classification positively.

**Effect of $K$** To understand the influence of the number of iterations on performance, we plot in Figure 1 the performance on ALL⁻ and ALL when increasing the $K$ hyperparameter from 0 to 20. As can be seen, the trend is quite clear: the model improves going from $K = 1$ to $K = 10$, but then increasing $K$ further results in (slightly) diminished performances. The reason for this could be that $K = 10$ strikes the best tradeoff between influence range increase and oversmoothing. In fact, the average shortest path between any two nodes in the WordNet graph is around 6.3.

**MFS/Unseen** So, where does the improvement when using our technique come from? To answer this question we have isolate two subsets of ALL⁻: a most frequent synset (MFS) set, including only instances in which one of the gold synsets associated with the instance is the MFS, and a much harder unseen one, in which the gold synsets never occur in the training set. We report the results in Table 3.

As can be seen the benefits of using $K = 10$ in our model are much stronger for the long tail of never occurring senses than for high-frequency senses. In fact, our $K = 10$ model trained on SemCor improves the performance of the $K = 1$ baseline on unseen synsets by 1.9 points (67.0 against 65.1), while the improvement on the MFS is much more modest (94.1 vs 93.9 F1).

## 5 Conclusion

In this paper we have shown that a deeper integration between supervised and knowledge-based methods in WSD can be attained. Indeed, by using the standard logits as personalization vector for an approximated neural PageRank, a supervised method can exploit not just *local* graph information, but *global* information as well. Thanks to our technique, we are able to match the best competitor system when training only on SemCor (Blevins and Zettlemoyer, 2020), while using a simpler model. When we concatenate the WNTG training corpus our results outperform the best competitor (Conia and Navigli, 2021) by 0.5 on the overall evaluation. We leave it as future work to ascertain whether i) edge label information can be incorporated too, and ii) stronger baseline models using glosses (Yap et al., 2020; Blevins and Zettlemoyer, 2020) can benefit from the use of our method.

## Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

## References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens, Greece. Association for Computational Linguistics.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

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 (EMNLP)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michele Bevilacqua and Roberto Navigli. 2020. 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, Online. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of Word Sense Disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2021. Framing Word Sense Disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275, Online. Association for Computational Linguistics.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France. Association for Computational Linguistics.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved Word Sense Disambiguation using pre-trained contextualized word representations. 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)*, pages 5297–5306, Hong Kong, China. Association for Computational Linguistics.

Taher H. Haveliwala. 2002. Topic-sensitive PageRank. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, page 517–526, New York, NY, USA. Association for Computing Machinery.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for Word Sense Disambiguation with gloss knowledge. 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)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. 2019. Combining neural networks with personalized PageRank for classification on graphs. In *Proceedings of the International Conference on Learning Representations*.

Daniel Loureiro and Alípio Jorge. 2019. 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*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. 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)*, pages 3534–3540, Hong Kong, China. Association for Computational Linguistics.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, pages 235–244.

George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. 1994. Using a semantic concordance for sense identification. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. 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)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017a. 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*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017b. Neural sequence learning models for Word Sense Disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. SensEmBERT: Context-enhanced sense embeddings for multilingual Word Sense Disambiguation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8758–8765. AAAI Press.

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual Word Sense Disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–46, Online. Association for Computational Linguistics.

Benjamin Snyder and Martha Palmer. 2004. 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*, pages 41–43, Barcelona, Spain. Association for Computational Linguistics.

Ming Wang and Yinglin Wang. 2021. Word sense disambiguation: Towards interactive context exploitation from both word and sense perspectives. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5218–5229, Online. Association for Computational Linguistics.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting BERT for Word Sense Disambiguation with gloss selection objective and example sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46, Online. Association for Computational Linguistics.