Biomedical Word Sense Disambiguation with Contextualized Representation Learning

Mozhgan Saeidi
mozhgan.saeidi@dal.ca
Dalhousie University
Canada

Evangelos Milios
eem@cs.dal.ca
Dalhousie University
Canada

Norbert Zeh
nzeh@cs.dal.ca
Dalhousie University
Canada

ABSTRACT

Representation learning is an important component in solving most Natural Language Processing (NLP) problems, including Word Sense Disambiguation (WSD). The WSD task tries to find the best meaning in a knowledge base for a word with multiple meanings (ambiguous word). WSD methods choose this best meaning based on the context, i.e., the words around the ambiguous word in the input text document. Thus, word representations may improve the effectiveness of the disambiguation models if they carry useful information from the context and the knowledge base. Most of the current representation learning approaches are that they are mostly trained on the general English text and are not domain specified. In this paper, we present a novel contextual-knowledge base aware sense representation method in the biomedical domain. The novelty in our representation is the integration of the knowledge base and the context. This representation lies in a space comparable to that of contextualized word vectors, thus allowing a word occurrence to be easily linked to its meaning by applying a simple nearest neighbor approach. Comparing our approach with state-of-the-art methods shows the effectiveness of our method in terms of text coherence.

KEYWORDS

Representation Learning, Neural Networks, Biomedical Text, Word Sense Disambiguation, Transformers

ACM Reference Format:
Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. 2022. Biomedical Word Sense Disambiguation with Contextualized Representation Learning. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, Lyon, France, 6 pages. https://doi.org/10.1145/3487553.3524703

1 INTRODUCTION

Natural Language Processing (NLP) includes many tasks which most of which are engaged with representation learning (RL) [8, 46]. Text representation learning has shown its important impact on the final results of NLP tasks, including Word Sense Disambiguation (WSD) [28], Information Retrieval (IR) [54], and Question Answering [51]. After the development of deep neural networks, different approaches have been widely used to solve the NLP tasks. Some of these deep neural networks are convolutional neural networks (CNNs) [17, 26, 55], recurrent neural networks (RNNs) [56], graph-based neural networks (GNNs) [35], and attention mechanisms [1]. Representation learning is also one of the tasks that uses the power of deep learning to alleviate feature engineering difficulties [45]. RL models usually use low-dimensional and dense vectors to implicitly represent the syntactic or semantic features of the language [44].

On the large corpus, the pre-trained models can learn language representations and then be used to solve downstream tasks. Between different RL approaches, the Skip-gram [24] and GloVe [30] are such models that are very shallow for computational efficiencies. While by emerging the deep models, including transformers [9], the RL architecture is transferred from shallow to deep. The pre-trained embeddings capture the semantics of the words they represent, but they suffer from the context in their representations [42]. The importance of the context in word representation is vital in some NLP tasks, like WSD.

The next generation of the pre-trained language models shows the important role of context in the representation in NLP tasks and tried enhancing RL with context [43]. Some of these models are including CoVe [22], ELMO [33], OpenAI, BERT [9], and GPT [37]. Most of the pre-trained language models are trained using English data sets. Because of this reason, their performance on different NLP tasks is good as long as the input text document is the general English text. If the text is in any specific domain, the representations are not that help solving the tasks.

The biomedical domain is one of these domains that needs specific pre-trained language models; this need is because of the volume of the biomedical text, which is growing with a good speed and needs analysis for different problems. On average, more than 3000 new articles are published every day in peer-reviewed journals, excluding pre-prints and technical reports such as clinical trial reports in various archives. Consequently, there is increasingly more demand for accurate biomedical text mining tools to extract text information. The deep learning models have been used in this domain in some of the NLP tasks.

To use the recent pre-trained language models in different tasks, we need to train them on biomedical text. Since most of them, like Word2Vec, ELMo and BERT are trained and tested mainly on datasets containing general domain texts (e.g. Wikipedia), it is difficult to estimate their performance on datasets containing biomedical texts. Previously, Word2Vec, one of the most widely known

WWW, April 25–29, 2022, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524703

context-independent word representation models, was trained on biomedical corpora, which contain terms and expressions usually not included in a general domain corpus. While ELMo and BERT have proven the effectiveness of contextualized word representations, they cannot obtain high performance on biomedical corpora because they are pre-trained on only general domain corpora. The other problem is the word distributions of general and biomedical corpora, which are quite different, and it can often be a problem for biomedical text mining models. As a result, recent models in biomedical text mining rely largely on adapted versions of word representations. The other recent works tried developing these pre-trained language models on biomedical texts and reduce this gap. These representations are still in lack of considering the context when generating representations, which it impacts the results on some tasks, like WSD.

In this study, first, we trained current state-of-the-art RL models on biomedical text. Second, we present a new representation learning approach (BioCBERT stands for Bio-Contextualized BERT) which considers the context of the input text and the context of the knowledge base when generating the embeddings. This information integration from context and the knowledge base is helpful in solving the NLP tasks with representations that carry context information with them. The context information in the word representations enhances the efficiency of various algorithms in NLP problems, including WSD. In the WSD problem, we try to find the best meaning for words that have multiple meanings. If the representations carry information from their input context, it enhances the WSD algorithm to find the best meaning match based on its context.

2 BACKGROUND

In this section, we first overview related works for the WSD task and then overview previous works toward pre-trained language models for lexical ambiguity. The WSD task is at the core of lexical semantics and has been tackled with many various approaches. We divide these approaches into two categories of knowledge-based and supervised approaches [28].

2.1 Knowledge-Based Approaches

Knowledge-based methods use the semantic network structure, e.g., Wikipedia [10], WordNet [25], or BabelNet [29], to find the correct meaning based on its context for each input word [27]. These approaches employ algorithms on graphs to address the word ambiguity in texts [2]. Disambiguation based on Wikipedia has been demonstrated to be comparable in terms of coverage to domain-specific ontology [53] since it has broad coverage, with documents about entities in a variety of domains [21]. The most widely used lexical knowledge base is WordNet, although it is restricted to the English lexicon, limiting its usefulness to other vocabularies. BabelNet solves this challenge by combining lexical and semantic information from various sources in numerous languages, allowing knowledge-based approaches to scale across all languages it supports. Despite their potential to scale across languages, knowledge-based techniques on English fall short of supervised systems in terms of accuracy [48]. One of the latest works in this series is SensEmBERT [48] which shows the power of language models combined with a vast amount of knowledge in a semantic network to produce latent semantic representations of nominal senses in multiple languages. ARES followed this model and created sense embeddings for the lexical meanings within a lexical knowledge base. These embeddings lie in a space that is comparable to that of contextualized word vectors [49].

2.2 Supervised Approaches

Supervised approaches use sense-annotated data for their training. These approaches surpass the knowledge-based ones in all English data sets, even before introducing pre-trained language models. These approaches use neural architectures [23], or SVM models [14], while still suffering from the need of creating large manually-curated corpora (knowledge acquisition bottleneck), which reduces their usability to scale over unseen words [11]. Automatic data augmentation approaches [47] developed methods to cover more words, senses, and languages.

Neural sequence models are trained for end-to-end WSD by Raganato et al. [38]. They re-framed WSD as a translation task that sequences of words are translated into sequences of senses. Later, some works showed the potential of contextual representation for WSD [31]. Sense embeddings initialization using glosses and adapted the skip-gram objective of word2vec is done by Chen et al. [6] to learn and improve the sense embeddings jointly with word embeddings. Later, by the appearance of NASARI vectors [5], sense embeddings were created using structural knowledge from a large multilingual semantic network. These methods represent sense embeddings in the same space as the pre-trained word embeddings, while they suffer from fixed embedding spaces. Finally, the LMMS representation considers creating sense-level embeddings with complete coverage of WordNet and shows the power of this representation for WSD by applying a simple Nearest Neighbors (k-NN) method [19]. ARES used this 1-NN method with its representations and showed improved results in the WSD task.

2.3 Language Modelling Representation

Most NLP tasks now use semantic representations derived from language models. There are static word embeddings and contextual embeddings.

2.3.1 Static Word Embeddings. Word embeddings are distributional semantic representations usually with one of two goals: predict context words given a target word (Skip-Gram), or the inverse (CBOW) [24]. In both, the target word is at the center, and the context is considered as a fixed-length window that slides over tokenized text. These models produce dense word representations. One limit for word embeddings means conflict around word types. This limitation affects the capability of these word embeddings for the ones that are sensitive to their context [40].

2.3.2 Contextual Word Embeddings. The problem mentioned as a limitation for the static word embeddings is solved in this type of embeddings. The critical difference is that the contextual embeddings are sensitive to the context. Therefore, it allows the same word types to have different representations according to their context. The first work in contextual embeddings is ELMO [31], which is followed by BERT [9], as the state-of-the-art model. The

critical feature of BERT, which makes it different, is the quality of its representations. Its results are task-specific fine-tuning of pre-trained neural language models. The recent representations which we analyze their effectiveness are based on this two models [34].

Transformer-based language models are pretty new in the NLP field, but there are a few works for analyzing these models and understanding the structure behind them [20]. The transformer-based models have been shown to capture the syntax and be applicable for solving the NLP problems [12]. Jawahar et al. [15] offers a phrasal representation analysis from BERT captured with the lower layers. It is also shown that transformer-based models encode well the human-like parse trees [13]. Quantitative analysis of contextualized word embeddings and sentence embedding models has demonstrated the effectiveness of the models' analysis of the semantic roles [31]. The role of models for encoding sentence structure across a range of syntactic, semantic, local, and long-range phenomena is examined by Tenney et al. [50] and shows the strength of representations for syntactic phenomena. The entity type exploration and their relations are described in [50]. The effectiveness of LSTM language models has been shown [18], as well as understanding their internal representations for predicting words in a context [52]. Furthermore, the LSTM predictions for a word in context provide the ability to retrieve substitutes, showing how well the language model has captured the information [4]. Finally, for this LSTM-based contextualized embedding model, some analyses show how well these models distinguish between usages of words in context [3].

In terms of a complete overview of neural network approaches and study of the BERT model, there are some complete recent surveys [41]. The geometry of BERT is quantified in Reif et al. [39] which shows how this model cares about the neighboring tokens. However, the role of language models in lexical ambiguity is not addressed in any of these works. A few studies try to use knowledge resources and extract semantic information to enhance the generalization of pre-trained language models like BERT [32]. Characterizing the sense representation of BERT using cluster analysis has also been studied [7]. The study on BERT's layers by Reif et al. [39] shows how this model performs for sense representations. The layer-wise performance of BERT when applied to the WSD task was studied in Loureiro et al. [20]. The difference of our research is to quantitatively understand to what extent the pre-trained language models encode information for the lexical ambiguity in terms of different word types. We show these pre-trained contextualized sense embedding behavior when solving the ambiguousness of part-of-speech in the text.

3 PRELIMINARIES

Our new BioCBERT (Biomedical Contextualized BERT embedding) representation uses different resources to build its vectors. In this section, we provide information on these resources.

BioBERT is a contextualized language representation model, based on BERT, a transformer-based language model for learning contextual representations of words in a text [9]. The contextualized representation of BERT is the key factor that has changed the performance in many NLP tasks, such as WSD. BioBERT is a pre-trained model which is trained on different combinations of general and biomedical domain corpora.

SBERT is a modification of the pre-trained BERT network that uses Siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. We use this sentence representation when generating the vector representations of sense sentences, both for the input text as the context and for the knowledge base text.

Wikipedia is a free content and multilingual online encyclopedia that uses a wiki-based editing system. Wikipedia is organized with Wikipedia pages, which are articles. Therefore, articles in Wikipedia can be directly linked to the entities they describe in other knowledge bases. Furthermore, mentions of entities in Wikipedia articles often provide a link to the relevant Wikipedia pages, thus providing labeled examples of mentions and associated anchor texts in various contexts, which could be used for supervised learning in WSD with Wikipedia as the knowledge base [57].

PubMed is a free search engine accessing the MEDLINE database of references and abstracts on life sciences and biomedical topics primarily. This preliminary is used at the time of generating BioBERT representations.

UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote the creation of more effective and interoperable biomedical information systems and services, including electronic health records.

4 BIOCBERT

The BioCBERT is a pre-trained language representation model for biomedical text. BioCBERT is created by combining semantic and textual information from the first paragraph of each sense's Wikipedia page and the paragraph of the input document text, which includes the senses. BioCBERT uses the representation power of neural language models, i.e., BioBERT and SBERT. The other preliminaries creating BioCBERT are Wikipedia, and UMLS. BioCBERT is based on three components; Context Retrieval, Word Embedding, and Sense Embedding.

4.1 Context Retrieval

This first component aims to collect contextual information from the knowledge base, which enhances the representations. For each ambiguous word in the input text, we create a set including candidate senses for the word from Wikipedia. This procedure aims to collect suitable contextual information from Wikipedia for each given concept in the semantic network. Then we exploit the mapping between synsets and Wikipedia pages available in the biomedical inventory and its taxonomic structure to collect textual information relevant to a target synset s. For each synset s, we collect all the connected concepts to s from the UMLS. We show this set of related synsets to s by R_s which is:

$$R_s = \{s' | (s, s') \in E\} \tag{1}$$

E is the set including all connections. In this work, for each page p_s, we consider the first opening paragraph of the page and compute its lexical vector by summing the SBERT vector representation of the sentences in this first paragraph. These lexical representations

are later used for the similarity score finding between p_s and $p_{s'}$, for each $s' \in R_s$ by using the weighted overlap measure from [36], which is defined as follows:

$$WO(p_1, p_2) = \left(\sum_{w \in O} \frac{1}{r_w^{p_1} + r_w^{p_2}}\right)\left(\sum_{i=1}^{|O|} \frac{1}{2i}\right)^{-1} \quad (2)$$

where O is the set of overlapping dimensions of p_1 and p_2 and $r_w^{p_i}$ is the rank of the word w in the lexical vector of p_i. We preferred the weighted overlap over the more common cosine similarity as it has proven to perform better when comparing sparse vector representations [36]. Once we have scored all the $(p_s, p_{s'})$ pairs, we create partitions of R_s, each comprising all the senses s' connected to s with the same relation r, where r can be one among connections. We then retain from each partition only the top-k scored senses according to $WO(p_{s_i}, p_{s'_i})$, which we set $k = 15$ in our experiments.

4.2 Word Embedding

In the second component, we use BioBERT to extract the given ambiguous word from the input text. For each ambiguous word (mention) of the input, we extract its BioBERT representation. Using the UMLS relations to Wikipedia, we extract all synsets of mention from UMLS (set E). For each of these senses, we collect all the Wikipedia pages for each sense. We use BioBERT representation for the second time to generate vector representation for senses.

4.3 Sense Embedding

In this last component, we build the final representation of each mention. From the previous step, we took the representation of mention, $R(m)$, and the representation of each one of its senses. We show the representations of each k sense of m by $R(s_i)$ which i varies from 1 to k. Our unique representations combine the mention representation with sense representation, concatenating the two vector representations of $R(m)$ and $R(s_i)$. If mention m has k senses, BioCBERT generates k different representations of $R(m, s_1)$, $R(m, s_2)$, …, $R(m, s_k)$. The next novelty in our BioCBERT representations is ranking the k senses of each mention based on their relevancy degree to the context. To this aim, we concatenate representations of the first step. In the first step, we took the representation of the input text paragraph, which contains the ambiguous mention, show it by $R(PD)$ which stands for representation of the **P**aragraph of the input **D**ocument text. In the first step, we also took the representation of the first paragraph of the Wikipedia page, which represents it by $R(PW)$, which stands for representation of the first **P**aragraph of the **W**ikipedia page. Finally, we concatenate these two representations as $R(PD, PW)$. The dimension of this concatenated representation is also equal to the word representation, making it possible to calculate their cosine similarities. To rank the senses relevancy's to the context, we use the cosine similarity as follows:

$$\text{Sim}(m, s_i) = Cosine(R(m, s_i), R(PD, PW)), \text{for } i = 1, \ldots, k \quad (3)$$

This ranking provides the most similar sense to the context for each mention. This novelty makes this representation more effective than the previous contextualized-based embeddings, especially in the task of word sense disambiguation. At the end of these three steps, each sense is associated with a vector that encodes both the contextual information and semantic knowledge base information from the extracted context of Wikipedia and its gloss.

5 EXPERIMENTAL SETUP

We present the settings of our evaluation of BioCBERT in the WSD task in biomedical text. This setup includes the benchmark, BioCBERT setup for disambiguation task and state-of-the-art WSD models as our comparison systems. To test each embedding on the WSD task, we employed the 1-NN algorithm and compared the disambiguated sense of each word with the ground truth annotations in the datasets. The nearest neighbors strategy is effective with pre-trained language models [20].

5.1 Evaluation Benchmark

The UMLS metathesaurus includes 3.4 million biomedical and clinical concepts. Each concept has a unique identifier called CUI (Concept Unique Identifier), a set of representative terms, and a text definition. The Metathesaurus provided us with the sense sets of ambiguous terms. The SPECIALIST Lexicon resource contains information about common English vocabulary and biomedical terms by offering tools for language processing. The next source is Medline which includes over 20 million citations of life sciences and biomedical articles from 1966 to the present. Combined with the UMLS concept definitions, we employed Medline 2013 bigram-list to create our sense embeddings. As validation datasets, we employed the MSH WSD [16] dataset for the evaluation of WSD algorithms[1]. This dataset provides 37888 instances for 203 ambiguous terms, including abbreviations, that take 2–5 senses (100 instances per each sense are provided). Prepared from Medline, every instance of a target ambiguous term is manually annotated with a CUI within the sense set of that term. For example, an instance of Ca is labeled with either C0006823 (Canada), C0006675 (California), C0006754 (calcium), or C3887642 (cornu ammonis); while every instance of the target term lymphogranulomatosis takes the sense C0036202 (benign lymphogranulomatosis) or C0019829 (malignant lymphogranulomatosis).

6 RESULTS

The results of our evaluations on the WSD task are represented in this section. We show the effectiveness of BioCBERT representation by comparing it with the existing state-of-the-art models. In Table 1 and Table 2 we report the results of BioCBERT and compare it against the results obtained from other state-of-the-art approaches. The performances are reported in terms of accuracy. As we can see, BioCBERT achieves the best results on the datasets compared to other previous contextualized approaches. It indicates that BioCBERT is competitive with these previous models. These results show that the novel idea in the nature of creating this BioCBERT representation has improved the lexical ambiguity. It is a good indicator of the dependency of the WSD task to the representation that is aware of the context and the information extracted from the reference knowledge base.

[1] https://lhncbc.nlm.nih.gov/ii/areas/WSD/collaboration.html

Table 1: Accuracy results for MSH WSD dataset with unsupervised methods.

Model	Macro-Accuracy	Micro-Accuracy
BioBERT	83.4	82.6
Bio Graph	71.52	-
deepBioWSD with random embeddings	92.16	91.93
deepBioWSD with pretrained embeddings	92.67	92.21
BioCBERT	94.71	93.84

Table 2: Accuracy results for MSH WSD dataset with supervised methods.

Model	Macro-Accuracy	Micro-Accuracy
NB	91.84	-
SVM	92.62	-
LSTM	91.87	91.78
BLSTM	93.64	92.47
BioCBERT	94.71	93.84

7 CONCLUSION

In this paper, we present BioCBERT, a novel approach for creating sense embeddings considering the knowledge base and the context of the biomedical input document text. We showed that this context-rich representation is beneficial for lexical ambiguity in the biomedical domain. The results of experiments in the WSD task show the efficiency of BioCBERT representations compared to other state-of-the-art methods, despite relying only on biomedical data. The results across other different datasets show the high quality of our embeddings and also enable the biomedical WSD while at the same time relieving the heavy requirement of sense-annotated corpora. We further tested our embeddings on the split data into four parts of speeches. By applying disambiguation on parts-of-speeches in the dataset, we show the efficiency of different representations. This work shows how context can play an important role in the final results of the word sense disambiguation systems. The integrated information from both the input text, as context, and from the knowledge base is leading the final choice of the disambiguation system as the correct meaning for the ambiguous words.

REFERENCES

[1] J. Uszkoreit L. Jones A. N. Gomez L. Kaiser A. Vaswani N. Shazeer N. Parmar and I. Polosukhin. 2017. Attention is all you need. *https://arxiv.org/abs/1706.03762* (2017). https://arxiv.org/abs/1706.03762

[2] Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics* 40, 1 (March 2014), 57–84. https://direct.mit.edu/coli/article/40/1/57/145

[3] Laura Aina, Kristina Gulordava, and Gemma Boleda. 2019. Putting words in context: LSTM language models and lexical ambiguity. *ACL* (2019), 3342–3348. https://www.aclweb.org/anthology/W19-0423/

[4] Asaf Amrami and Yoav Goldberg. 2018. Word sense induction with neural biLM and symmetric patterns. *In Proceeding of EMNLP* (2018), 4860–4867. https://www.aclweb.org/anthology/D18-1523/

[5] Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Artificial Intelligence Research* 63 (2018), 743–788.

[6] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference EMNLP*. 1025–1035. https://www.aclweb.org/anthology/D14-1110/

[7] Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference CNLL*. 227–244. https://www.aclweb.org/anthology/2020.conll-1.17/

[8] Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. *arXiv preprint arXiv:1902.10547* (2019).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL* (2018). https://www.aclweb.org/anthology/N19-1423/

[10] Angela Fogarolli. 2009. Word sense disambiguation based on wikipedia link structure. In *2009 IEEE International Conference on Semantic Computing*. IEEE, 77–82. https://ieeexplore.ieee.org/stamp/stamp.jsp

[11] William A Gale, Kenneth W Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26, 5 (1992), 415–439.

[12] Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *CoRR, abs/1901.05287* (2019). https://arxiv.org/abs/1901.05287

[13] John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *ACL*. 4129–4138. https://www.aclweb.org/anthology/N19-1419/

[14] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *ACL*. 897–907. https://www.aclweb.org/anthology/P16-1085.pdf

[15] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language?. In *ACL*. https://www.aclweb.org/anthology/P19-1356/

[16] Antonio J Jimeno-Yepes, Bridget T McInnes, and Alan R Aronson. 2011. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC bioinformatics* 12, 1 (2011), 1–14.

[17] Tuan Lai, Heng Ji, and ChengXiang Zhai. 2021. BERT might be Overkill: A Tiny but Effective Biomedical Entity Linker based on Residual Convolutional Neural Networks. *arXiv preprint arXiv:2109.02237* (2021).

[18] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4 (2016), 521–535. https://www.aclweb.org/anthology/Q16-1037/

[19] Daniel Loureiro and Alipio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. *ACL* (2019), 5682–5691. https://www.aclweb.org/anthology/P19-1569

[20] Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics* (2021), 1–55. https://direct.mit.edu/coli/article/doi/10.1162/coli_a_00405/98520/Analysis-and-Evaluation-of-Language-Models-for

[21] Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. 2020. Information extraction meets the semantic web: a survey. *Semantic Web* Preprint (2020), 1–81. http://repositorio.uchile.cl/bitstream/handle/2250/174484/Information-extraction-meets-the-Semantic-Web.pdf?sequence=1

[22] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *arXiv preprint arXiv:1708.00107* (2017).

[23] Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *SIGNLL*. 51–61. https://www.aclweb.org/anthology/K16-1006/

[24] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of ICLR* 4 (2013), 321–329. https://arxiv.org/pdf/1301.3781.pdf

[25] 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* 3, 4 (1990), 235–244. https://watermark.silverchair.com/235.pdf

[26] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep Learning–based Text Classification: A Comprehensive Review. *ACM Computing Surveys (CSUR)* 54, 3 (2021), 1–40.

[27] 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 (2014), 231–244. https://watermark.silverchair.com/tacl_a_00179.pdf

[28] Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)* 41, 2 (2009), 1–69. https://dl.acm.org/doi/abs/10.1145/1459352.1459355

[29] Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence* 193 (2012), 217–250. https://www.sciencedirect.com/science/article/pii/S0004370212000793

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. EMNLP, Qatar, 1532–1543. https://www.aclweb.org/anthology/D14-1162.pdf

[31] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. *ACL* (2018), 2227–2237. https://www.aclweb.org/anthology/N18-1202

[32] Matthew E Peters, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164* (2019). https://arxiv.org/pdf/1909.04164.pdf

[33] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

[34] Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. *EMNLP* (2018), 1499–1509. https://www.aclweb.org/anthology/D18-1179/

[35] Minh Tran Phu and Thien Huu Nguyen. 2021. Graph Convolutional Networks for Event Causality Identification with Rich Document-level Structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3480–3490.

[36] Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *ACL*. 1341–1351. https://www.aclweb.org/anthology/P13-1132.pdf

[37] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[38] Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *EMNLP*. 1156–1167.

[39] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems* 32 (2019), 8594–8603. https://proceedings.neurips.cc/paper/2019/hash/159c1ffe5b61b41b3c4d8f4c2150f6c4-Abstract.html

[40] Joseph Reisinger and Raymond Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: ACL*. 109–117. https://www.aclweb.org/anthology/N10-1013.pdf

[41] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics* 8 (2020), 842–866. https://www.aclweb.org/anthology/2020.tacl-1.54/

[42] Mozhgan Saeidi. 2021. ContextBERT: Contextual Graph Representation Learning in Text Disambiguation. (2021), 283–297. http://ceur-ws.org/Vol-2997/paper2.pdf

[43] Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. 2021. Contextualized Knowledge Base Sense Embeddings in Word Sense Disambiguation. In *International Conference on Document Analysis and Recognition*. Springer, 174–186.

[44] Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. 2021. Graph Convolutional Networks for Categorizing Online Harassment on Twitter. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 946–951.

[45] Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. 2021. Graph representation learning in document wikification. In *International Conference on Document Analysis and Recognition*. Springer, 509–524.

[46] Mozhgan Saeidi, Samuel Bruno da S Sousa, Evangelos Milios, Norbert Zeh, and Lilian Berton. 2019. Categorizing online harassment on Twitter. In *Joint European Conference on Machine Learning and KDD*. Springer, 283–297. https://link.springer.com/chapter/10.1007/978-3-030-43887-6_22

[47] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. Just "OneSeC" for producing multilingual sense-annotated data. In *ACL*. 699–709. https://www.aclweb.org/anthology/P19-1069.pdf

[48] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *AAAI*, Vol. 34. 8758–8765. https://ojs.aaai.org//index.php/AAAI/article/view/6402

[49] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round Word Sense Disambiguation. In *EMNLP*. 3528–3539. https://www.aclweb.org/anthology/2020.emnlp-main.285/

[50] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *In International Conference on Learning Representations* (2019). https://arxiv.org/abs/1905.06316

[51] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 355–363.

[52] Marten Van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. *EMNLP-IJCNLP* (2019), 5831–5837. https://www.aclweb.org/anthology/D19-1592/

[53] Gerhard Weikum, Luna Dong, Simon Razniewski, and Fabian Suchanek. 2020. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. *arXiv preprint arXiv:2009.11564* (2020). https://arxiv.org/pdf/2009.11564.pdf

[54] Robert West, Ashwin Paranjape, and Jure Leskovec. 2015. Mining missing hyperlinks from human navigation traces: A case study of Wikipedia. In *Proceedings of the 24th international conference on World Wide Web*. 1242–1252. https://dl.acm.org/doi/pdf/10.1145/2736277.2741666

[55] Jin-Wen Wu, Fei Yin, Yan-Ming Zhang, Xu-Yao Zhang, and Cheng-Lin Liu. 2021. Graph-to-Graph: Towards Accurate and Interpretable Online Handwritten Mathematical Expression Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2925–2933.

[56] Jian Zhang, Ke Yan, and Yuchang Mo. 2021. Multi-Task Learning for Sentiment Analysis with Hard-Sharing and Task Recognition Mechanisms. *Information* 12, 5 (2021), 207.

[57] Gang Zhao, Ji Wu, Dingding Wang, and Tao Li. 2016. Entity disambiguation to Wikipedia using collective ranking. *Information Processing & Management* 52, 6 (2016), 1247–1257. https://www.sciencedirect.com/science/article/pii/S0306457316301893