

Does BERT Make Any Sense?

Interpretable Word Sense Disambiguation with Contextualized Embeddings

Gregor Wiedemann¹ Steffen Remus¹ Avi Chawla² Chris Biemann¹

¹Language Technology Group
Department of Informatics

Universität Hamburg, Germany

{gwiedemann, remus, biemann}
@informatik.uni-hamburg.de

²Indian Institute of Technology (BHU)
Varanasi, India

avi.chawla.cse16@iitbhu.ac.in

Abstract

Contextualized word embeddings (CWE) such as provided by ELMo (Peters et al., 2018), Flair NLP (Akbik et al., 2018), or BERT (Devlin et al., 2019) are a major recent innovation in NLP. CWEs provide semantic vector representations of words depending on their respective context. Their advantage over static word embeddings has been shown for a number of tasks, such as text classification, sequence tagging, or machine translation. Since vectors of the same word type can vary depending on the respective context, they implicitly provide a model for word sense disambiguation (WSD). We introduce a simple but effective approach to WSD using a nearest neighbor classification on CWEs. We compare the performance of different CWE models for the task and can report improvements above the current state of the art for two standard WSD benchmark datasets. We further show that the pre-trained BERT model is able to place polysemic words into distinct ‘sense’ regions of the embedding space, while ELMo and Flair NLP do not seem to possess this ability.

1 Synonymy and Polysemy of Word Representations

Lexical semantics is characterized by a high degree of polysemy, i.e. the meaning of a word changes depending on the context in which it is currently used (Harris, 1954). Word Sense Disambiguation (WSD) is the task to identify the correct sense of the usage of a word from a (usually) fixed inventory of sense identifiers. For the English language, WordNet (Fellbaum, 1998) is the most commonly used sense inventory providing more than 200K word-sense pairs.

To train and evaluate WSD systems, a number of shared task datasets have been published in the SemEval workshop series. In the *lexical sample* task (Kilgarriff, 2001; Mihalcea et al., 2004), a training set and a test set is provided. The relatively large data contains one sense-annotated word per training/test instance. The *all-words* task (Edmonds and Cotton, 2001; Snyder and Palmer, 2004) only provides a small number of documents as test data where each ambiguous word is annotated with its sense. To facilitate the comparison of WSD systems, some efforts have been made to provide a comprehensive evaluation framework (Raganato et al., 2017), and to unify all publicly available datasets for the English language (Vial et al., 2018b).

WSD systems can be distinguished into three types — knowledge-based, supervised, and semi-supervised approaches. *Knowledge-based* systems utilize language resources such as dictionaries, thesauri and knowledge graphs to infer senses. *Supervised* approaches train a machine classifier to predict a sense given the target word and its context based on an annotated training data set. *Semi-supervised* approaches extend manually created training sets by large corpora of unlabeled data to improve WSD performance. All approaches rely on some way of context representation to predict the correct sense. Context is typically modeled via dictionary resources linked with senses, or as some feature vector obtained from a machine learning model.

A fundamental assumption in structuralist linguistics is the distinction between signifier and signified as introduced by Ferdinand de Saussure (Saussure, 2001) in the early 20th century. Computational linguistic approaches, when using character strings as the only representatives for word meaning, implicitly assume identity between signifier and signified. Different word senses are simply collapsed into the same string representation. In

this respect, word counting and dictionary-based approaches to analyze natural language texts have been criticized as pre-Saussurean (Pêcheux et al., 1995). In contrast, the distributional hypothesis not only states that meaning is dependent on context. It also states that words occurring in the same contexts tend to have a similar meaning (Harris, 1954). Hence, a more elegant way of representing meaning has been introduced by using the contexts of a word as an intermediate semantic representation that mediates between signifier and signified. For this, explicit vector representations, such as TF-IDF, or latent vector representations, with reduced dimensionality, have been widely used. Latent vector representations of words are commonly called word embeddings. They are fixed length vector representations, which are supposed to encode semantic properties. The seminal neural word embedding model Word2Vec (Mikolov et al., 2013), for instance, can be trained efficiently on billions of sentence contexts to obtain semantic vectors, one for each word type in the vocabulary. It allows synonymous terms to have similar vector representations that can be used for modeling virtually any downstream NLP task. Still, a polysemic term is represented by one single vector only, which represents all of its different senses in a collapsed fashion.

To capture polysemy as well, the idea of word embeddings has been extended to encode word sense embeddings. Neelakantan et al. (2014) first introduced a neural model to learn multiple embeddings for one word depending on different senses. The number of senses can be defined by a given parameter, or derived automatically in a non-parametric version of the model. However, employing sense embeddings in any downstream NLP task requires a reliable WSD system in an earlier stage to decide how to choose the appropriate embedding from the sense inventory.

Recent efforts to capture polysemy for word embeddings give up on the idea of a fixed word sense inventory. Contextualized word embeddings (CWE) do not only create one vector representation for each type in the vocabulary, they also they produce distinct vectors for each token in a given context. The contextualized vector representation is supposed to represent word meaning and context information. This enables downstream tasks to actually distinguish the two levels of the signifier and the signified allowing for more realistic modeling

of natural language. The advantage of such contextually embedded token representations compared to static word embeddings has been shown for a number of tasks such as text classification (Zampieri et al., 2019) and sequence tagging (Akbik et al., 2018).

Contribution: We show that CWEs can be utilized directly to approach the WSD task due to their nature of providing distinct vector representations for the same token depending on its context. To learn the semantic capabilities of CWEs, we employ a simple, yet interpretable approach to WSD using a k -nearest neighbor classification (kNN) approach. We compare the performance of three different CWE models on four standard benchmark datasets. Our evaluation yields that not all contextualization approaches are equally effective in dealing with polysemy, and that the simple kNN approach suffers severely from sparsity in training datasets. Yet, by using kNN, we include provenance into our model, which allows to investigate the training sentences that lead to the classifier’s decision. Thus, we are able to study to what extent polysemy is captured by a specific contextualization approach. For two datasets, we are able to report new state-of-the-art (SOTA) results.

2 Related Work

2.1 Neural Word Sense Disambiguation

Several efforts have been made to induce different vectors for the multiplicity of senses a word can express. Bartunov et al. (2016), Neelakantan et al. (2014), or Rothe and Schütze (2015) induce so-called sense embeddings in a pre-training fashion. While Bartunov et al. (2016) induce sense embeddings in an unsupervised way and only fix the maximum number of senses per word, Rothe and Schütze (2015) require a pre-labeled sense inventory such as WordNet. Then, the sense embeddings are mapped to their corresponding synsets. Other approaches include the re-use of pre-trained word embeddings in order to induce new sense embeddings (Pevina et al., 2016; Remus and Biemann, 2018). Panchenko et al. (2017) then also use induced sense embeddings for the downstream task of WSD. Camacho-Collados and Pilehvar (2018) provide an extensive overview of different word sense modeling approaches.

For WSD, (semi-)supervised approaches with recurrent neural network architectures represent the

current state of the art. Two major approaches were followed. First, Melamud et al. (2016) and Yuan et al. (2016), for instance, compute sentence context vectors for ambiguous target words. In the prediction phase, they select nearest neighbors of context vectors to determine the target word sense. Yuan et al. (2016) also use unlabeled sentences in a semi-supervised label propagation approach to overcome the sparse training data problem of the WSD task. Second, Kågebäck and Salomonsson (2016) employ a recurrent neural network to classify sense labels for an ambiguous target word given its surrounding sentence context. In contrast to earlier approaches, which relied on feature engineering (Taghipour and Ng, 2015), their architecture only uses pretrained GloVe word embeddings (Pennington et al., 2014) to achieve SOTA results on two English lexical sample datasets. For the all-words WSD task, Vial et al. (2018a) also employ a recurrent neural network. But instead of single target words, they sequentially classify sense labels for all tokens in a sentence. They also introduce an approach to collapse the sense vocabulary from WordNet to unambiguous hypersenses, which increases the label to sample ratio for each label, i.e. sense identifier. By training their network on the large sense annotated datasets SemCor (Miller et al., 1993) and the Princeton Annotated Gloss Corpus based on WordNet synset definitions (Fellbaum, 1998), they achieve the highest performance so far on most all-words WSD benchmarks. A similar architecture with an enhanced sense vocabulary compression was applied in (Vial et al., 2019), but instead of GloVe embeddings, BERT wordpiece embeddings (Devlin et al., 2019) are used as input for training. Especially the BERT embeddings further improved the performance yielding new state-of-the-art results.

2.2 Contextualized Word Embeddings

The idea of modeling sentence or context-level semantics together with word-level semantics proved to be a powerful innovation. For most downstream NLP tasks, CWEs drastically improved the performance of neural architectures compared to static word embeddings. However, the contextualization methodologies differ widely. We, thus, hypothesize that they are also very different in their ability to capture polysemy.

Like static word embeddings, CWEs are trained on large amounts of unlabeled data by some vari-

ant of language modeling. In our study, we investigate three most prominent and widely applied approaches: **Flair** (Akbik et al., 2018), **ELMo** (Peters et al., 2018), and **BERT** (Devlin et al., 2019).

Flair: For the contextualization provided in the *Flair NLP* framework, Akbik et al. (2018) take a static pre-trained word embedding vector, e.g. the GloVe word embeddings (Pennington et al., 2014), and concatenate two context vectors based on the left and right sentence context of the word to it. Context vectors are computed by two recurrent neural models, one character language model trained from left to right, one another from right to left. Their approach has been applied successfully especially for sequence tagging tasks such as named entity recognition and part-of-speech tagging.

ELMo: *Embeddings from language models* (ELMo) (Peters et al., 2018) approaches contextualization similar to Flair, but instead of two character language models, two stacked recurrent models for words are trained, again one left to right, and another right to left. For CWEs, outputs from the embedding layer, and the two bidirectional recurrent layers are not concatenated, but collapsed into one layer by a weighted, element-wise summation.

BERT: In contrast to the previous two approaches, *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019) does not rely on the merging of two uni-directional recurrent language models with a (static) word embedding, but provides contextualized token embeddings in an end-to-end language model architecture. For this, a self-attention based transformer architecture is used, which, in combination with a masked language modeling target, allows to train the model seeing all left and right contexts of a target word at the same time. Self-attention and non-directionality of the language modeling task result in extraordinary performance gains compared to previous approaches.

According to the distributional hypothesis, if the same word regularly occurs in different, distinct contexts, we may assume polysemy of its meaning (Miller and Charles, 1991). Contextualized embeddings should be able to capture this property. In the following experiments, we investigate this hypothesis on the example of the introduced models.

	SE-2 (Tr)	SE-2 (Te)	SE-3 (Tr)	SE-3 (Te)	S7-T7 (coarse)	S7-T17 (fine)	SemCor	WNGT
#sentences	8,611	4,328	7,860	3,944	126	245	37,176	117,659
#CWEs	8,742	4,385	9,280	4,520	455	6,118	230,558	1,126,459
#distinct words	313	233	57	57	327	1,177	20,589	147,306
#senses	783	620	285	260	371	3,054	33,732	206,941
avg #senses p. word	2.50	2.66	5.00	4.56	1.13	2.59	1.64	1.40
avg #CWEs p. word & sense	11.16	7.07	32.56	17.38	1.23	2.00	6.83	5.44
avg k'	2.75	-	7.63	-	-	-	3.16	2.98

Table 1: Properties of our datasets. For the test sets (Te), we do not report k' since they are not used as kNN training instances.

3 Nearest Neighbor Classification for WSD

We employ a rather simple approach to WSD using non-parametric nearest neighbor classification (kNN) to investigate the semantic capabilities of contextualized word embeddings. Compared to parametric classification approaches such as support vector machines or neural models, kNN has the advantage that we can directly investigate the training examples that lead to a certain classifier decision.

The kNN classification algorithm (Cover and Hart, 1967) assigns a plurality vote of a sample’s nearest labeled neighbors in its vicinity. In the most simple case, one-nearest neighbor, it predicts the label from the nearest training instance by some defined distance metric. Although complex weighting schemes for kNN exist, we stick to the simple non-parametric version of the algorithm to be able to better investigate the semantic properties of different contextualized embedding approaches.

As distance measure for kNN, we rely on cosine distance of the CWE vectors. Our approach considers only senses for a target word that have been observed during training. We call this approach localized nearest neighbor word sense disambiguation. We use spaCy¹ (Honnibal and Johnson, 2015) for pre-processing and the lemma of a word as the target word representation, e.g. ‘danced’, ‘dances’ and ‘dancing’ are mapped to the same lemma ‘dance’. Since BERT uses wordpieces, i.e. subword units of words instead of entire words or lemmas, we re-tokenize the lemmatized sentence and average all wordpiece CWEs that belong to the target word. Moreover, for the experiments with BERT embeddings², we follow the heuristic by Devlin et al. (2019) and concatenate the averaged wordpiece vectors of the last four layers.

¹<https://spacy.io/>

²We use the bert-large-uncased model.

We test different values for our single hyperparameter $k \in \{1, \dots, 10, 50, 100, 500, 1000\}$. Like words in natural language, word senses follow a power-law distribution. Due to this, simple baseline approaches for WSD like the *most frequent sense* (MFS) baseline are rather high and hard to beat. Another effect of the skewed distribution are imbalanced training sets. Many senses described in WordNet only have one or two example sentences in the training sets, or are not present at all. This is severely problematic for larger k and the default implementation of kNN because of the majority class dominating the classification result. To deal with sense distribution imbalance, we modify the majority voting of kNN to $k' = \min(k, |V_s|)$ where V_s is the set of CWEs with the least frequent training examples for a given word sense s .

4 Datasets

We conduct our experiments with the help of four standard WSD evaluation sets, two lexical sample tasks and two all-words tasks. As lexical sample tasks, SensEval-2 (Kilgarriff, 2001, SE-2) and SensEval-3 (Mihalcea et al., 2004, SE-3) provide a training data set and test set each. The all-words tasks of SemEval 2007 Task 7 (Navigli et al., 2007, S7-T7) and Task 17 (Pradhan et al., 2007, S7-T17) solely comprise test data, both with a substantial overlap of their documents. The two sets differ in granularity: While ambiguous terms in Task 17 are annotated with one WordNet sense only, in Task 7 annotations are coarser clusters of highly similar WordNet senses. For training of the all-words tasks, we use *a*) the SemCor dataset (Miller et al., 1993), and *b*) the Princeton WordNet gloss corpus (WNGT) (Fellbaum, 1998) separately to investigate the influence of different training sets on our approach. For all experiments, we utilize the suggested datasets as provided by the UFSAC

Model	SE-2	SE-3	S7-T7 (coarse)		S7-T17 (fine)	
			SemCor	WNGT	SemCor	WNGT
Flair	65.27	68.75	69.24	78.68	45.92	50.99
ELMo	67.57	70.70	70.80	79.12	52.61	50.11
BERT	76.10	78.62	73.61	81.11	59.82	55.16

Table 2: kNN with $k = 1$ WSD performance (F1%). Best results for each testset are marked bold.

framework³ (Vial et al., 2018b), i.e. the respective training data. A concise overview of the data can be found in Table 1. From this, we can observe that the SE-2 and SE-3 training data sets, which were published along with the respective test sets, provide many more examples per word and sense than SemCor or WNGT.

5 Experimental Results

We conduct two experiments to determine whether contextualized word embeddings can solve the WSD task. In our first experiment, we compare different pre-trained embeddings with $k = 1$. In our second experiment, we test multiple values of k and the BERT pre-trained embeddings⁴ in order to estimate an optimal k . Further, we qualitatively examine the results to analyze, which cases can be typically solved by our approach and where it fails.

5.1 Contextualized Embeddings

To compare different CWE approaches, we use $k = 1$ nearest neighbor classification. Table 2 shows a high variance in performance. Simple kNN with ELMo as well as BERT embeddings beats the state of the art of the lexical sample task SE-2 (cp. Table 3). BERT also outperforms all others on the SE-3 task.

However, we observe a major performance drop of our approach for the two all-words WSD tasks in which no training data is provided along with the test set. For S7-T7 and S7-T17, the content and structure of the out-of-domain SemCor and WNGT training datasets differ drastically from those in the test data, which prevents yielding near state-of-the-art results. In fact, similarity of contextualized embeddings largely relies on semantically *and* structurally similar sentence contexts of polysemic target words. Hence, the more example sentences can be used for a sense, the higher are the chances

³Unification of Sense Annotated Corpora and Tools. We are using Version 2.1: <https://github.com/getalp/UFSC>

⁴BERT performed best in experiment one.

k	SE-2	SE-3	S7-T7	S7-T17
1	76.10	78.62	81.11	59.82
2	76.10	78.62	81.11	59.82
3	<u>76.52</u>	79.66	80.94	59.38
4	<u>76.52</u>	79.55	80.94	59.82
5	76.43	79.79	81.07	60.27
6	76.43	79.81	81.07	60.27
7	76.50	80.02	81.03	60.49
8	76.50	79.86	81.03	60.49
9	76.40	79.97	81.03	60.49
10	76.40	80.12	81.03	60.49
50	76.43	79.66	81.11	<u>60.94</u>
100	76.43	79.63	<u>81.20</u>	60.71
500	76.38	79.63	81.11	60.71
1000	76.38	79.63	81.11	60.71
MFS	54.79	58.95	70.94	48.44
Kågebäck (2016)	<i>66.90</i>	<i>73.40</i>	-	-
Yuan et al. (2016)	-	-	84.30	63.70
Vial et al. (2018a)	-	-	86.02	66.81
Vial et al. (2019)	-	-	90.60	71.40

Table 3: Best kNN models vs. most frequent sense (MFS) and state of the art (F1%). Best results are bold, previous SOTA is in italics and our best results are underlined.

that a nearest neighbor expresses the same sense. As can be seen in Table 1, the SE-2 and SE-3 training datasets provide more CWEs for each word and sense, and our approach performs better with a growing number of CWEs, even with a higher average number of senses per word as is the case in SE-3.

Thus, we conclude that the nearest neighbor approach suffers specifically from data sparseness. The chances increase that aspects of similarity other than the sense of the target word in two compared sentence contexts drive the kNN decision. Moreover, CWEs actually do not organize well in spherically shaped form in the embedding space. Although senses might be actually separable, the non-parametric kNN classification is unable to learn a complex decision boundary focusing only on the most informative aspects of the CWE (Yuan et al., 2016, p. 4).

5.2 Nearest Neighbors

K-Optimization: To attenuate for noise in the training data, we optimize for k to obtain a more robust nearest neighbor classification. Table 3 shows our best results using the BERT embeddings along with results from related works. For SensEval-2

and SensEval-3, we achieve a new state-of-the-art result. We observe convergence with higher k values since our k' normalization heuristic is activated. For the S7-T*, we also achieve minor improvements with a higher k , but still drastically lack behind the state of the art.

Senses in CWE space: We investigate how well different CWE models encode information such as distinguishable senses in their vector space. Figure 1 shows T-SNE plots (van der Maaten and Hinton, 2008) of six different senses of the word “bank” in the SE-3 training dataset encoded by the three different CWE methods. For visualization purposes, we exclude senses with a frequency of less than two. The Flair embeddings hardly allow to distinguish any clusters as most senses are scattered across the entire plot. In the ELMo embedding space, the major senses are slightly more separated in different regions of the point cloud. Only in the BERT embedding space, some senses form clearly separable clusters. Also within the larger clusters, single senses are spread mostly in separated regions of the cluster. Hence, we conclude that BERT embeddings actually seem to encode some form of sense knowledge, which also explains why kNN can be successfully applied to them. Moreover, we can see why a more powerful parametric classification approach such as employed by Vial et al. (2019) is able to learn clear decision boundaries. Such clear decision boundaries seem to successfully solve the data sparseness issue of kNN.

Error analysis: From a qualitative inspection of true positive and false positive predictions, we are able to infer some semantic properties of the BERT embedding space and the used training corpora. Table 4 shows selected examples of polysemic words in different test sets, including their nearest neighbor from the respective training set.

Not only vocabulary overlap in the context as in ‘along the *bank* of the river’ and ‘along the *bank* of the river Greta’ (2) allows for correct predictions, but also semantic overlap as in ‘little earthy bank’ and ‘huge bank [of snow]’ (3). On the other hand, vocabulary overlap, as well as semantic relatedness as in ‘land bank’ (5) can lead to false predictions. Another interesting example for the latter is the confusion between ‘grass bank’ and ‘river bank’ (6) where the nearest neighbor sentence in the training set shares some military context with the target sentence. The correct sense (*bank%1:17:01::Sloping*

Land) and the predicted sense (*bank%1:17:00::A Long Ridge or Pile [of earth]*) share high semantic similarity, too. In this example, they might even be used interchangeably. Apparently this context yields higher similarity than any of the other training sentences containing ‘grass bank’ explicitly.

In Example (10), the targeted sense is an action, i.e. a verb sense, while the predicted sense is a noun, i.e. a different word class. In general, this could be easily fixed by restricting the classifier decision to the desired POS. However, while example (12) is still a false positive, it nicely shows that the simple kNN approach is able to distinguish senses by word class even though BERT never learned POS classes explicitly. This effect has been investigated in-depth by Jawahar et al. (2019), who found that each BERT layer learns different structural aspects of natural language. Example (12) also emphasizes the difficulty of distinguishing verb senses itself, i.e. the correct sense label in this example is *watch%2:39:00::look attentively* whereas the predicted label and the nearest neighbor is *watch%2:41:00::follow with the eyes or the mind; observe*. Verb senses in WordNet are very fine grained and thus harder to distinguish automatically and by humans, too.

5.3 Post-evaluation experiment

In order to address the issue of mixed POS senses, we run a further experiment, which restricts words to their lemma and their POS tag. Table 5 shows that including the POS restriction increases the F1 scores for S7-T7 and S7-T17. This can be explained by the number of different POS tags that can be found in the different corpora (c.f. Table 6). The results are more stable with respect to their relative performance, i.e. SemCor and WNGT reach comparable scores on S7-T17. Also, the results for SE-2 and SE-3 did not change drastically. This can be explained by the average number of POS tags a certain word is labeled with. This variety is much stronger in the S7-T* tasks compared to SE-*

6 Conclusion

In this paper, we tested the semantic properties of contextualized word embeddings (CWEs) to address word sense disambiguation.⁵ To test their capabilities to distinguish different senses of a particular word, by placing their contextualized vector

⁵The source code of our experiments is publicly available at: <https://github.com/uhh-lt/bert-sense>

Example sentence		Nearest neighbor	
SE-3 (train)		SE-3 (test)	
(1)	President Aquino, admitting that the death of Ferdinand Marcos had sparked a wave of sympathy for the late dictator, urged Filipinos to stop weeping for the man who had laughed all the way to the bank _[A Bank Building] .	They must have been filled in at the bank _[A Bank Building] either by Mr Hatton himself or else by the cashier who was attending to him.	
(2)	Soon after setting off we came to a forested valley along the banks _[Sloping Land] of the Gwaun.	In my own garden the twisted hazel, corylus avellana contorta, is underplanted with primroses, bluebells and wood anemones, for that is how I remember them growing, as they still do, along the banks _[Sloping Land] of the rive Greta	
(3)	In one direction only a little earthy bank _[A Long Ridge] separates me from the edge of the ocean, while in the other the valley goes back for miles and miles.	The lake has been swept clean of snow by the wind, the sweepings making a huge bank _[A Long Ridge] on our side that we have to negotiate.	
(4)	However, it can be possible for the documents to be signed after you have sent a payment by cheque provided that you arrange for us to hold the cheque and not pay it into the bank _[A Financial Institution] until we have received the signed deed of covenant.	The purpose of these stubs in a paying – in book is for the holder to have a record of the amount of money he had deposited in his bank _[A Bank Building] .	
(5)	He continued: assuming current market conditions do not deteriorate further, the group, with conservative borrowings, a prime land bank _[A Financial Institution] and a good forward sales position can look forward to another year of growth.	Crest Nicholson be the exception, not have much of a land bank _[Supply or Stock] and rely on its skill in land buying.	
(6)	The marine said, get down behind that grass bank _[A Long Ridge] , sir, and he immediately lobbed a mills grenade into the river.	The guns were all along the river bank _[Sloping Land] as far as I could see.	
SemCor		S7-T17	
(7)	Some 30 balloon _[Large Tough Nonrigid Bag] shows are held annually in the U.S., including the world’s largest convocation of ersatz Phineas Foggs – the nine-day Albuquerque International Balloon Fiesta that attracts some 800,000 enthusiasts and more than 500 balloons, some of which are fetchingly shaped to resemble Carmen Miranda, Garfield or a 12-story-high condom.	Homes and factories and schools and a big wide federal highway, instead of peaceful corn to rest your eyes on while you tried to rest your heart, while you tried not to look at the balloon _[Large Tough Nonrigid Bag] and the bandstand and the uniforms and the flash of the instruments.	
(8)	The condom balloon _[Large Tough Nonrigid Bag] was denied official entry status this year.	Just like the balloon _[Large Tough Nonrigid Bag] would go up and you could sit all day and wish it would spring a leak or blow to hell up and burn and nothing like that would happen.	
(9)	Starting with congressman Mario Biaggi (now serving a jail sentence _[The Period of Time a Prisoner Is Imprisoned]), the company began a career of bribing federal, state and local public officials and those close to public officials, right up to and including E. Robert Wallach, close friend and adviser to former attorney general Ed Meese.	When authorities convicted him of practicing medicine without a license (he got off with a suspended sentence _[The Period of Time a Prisoner Is Imprisoned] of three years because of his advanced age of 77), one of his victims was not around to testify: he was dead of cancer.”	
(10)	Americans it seems have followed Malcolm Forbes’s hot-air lead and taken to balloon _[To Ride in a Hot-Air Balloon] in a heady way.	Just like the balloon _[Large Tough Nonrigid Bag] would go up and you could sit all day and wish it would spring a leak or blow to hell up and burn and nothing like that would happen.	
(11)	Any question as to why an author would believe this plaintive, high-minded note of assurance is necessary is answered by reading this book _[A Published Written Work] about sticky fingers and sweaty scammers.	But the book _[A Written Version of a Play] is written around a somewhat dizzy cartoonist, and it has to be that way.	
(12)	In between came lots of coffee drinking while watching _[To Look Attentively] the balloons inflate and lots of standing around deciding who would fly in what balloon and in what order [...].	So Captain Jenks returned to his harbor post to watch _[To Follow With the Eyes or the Mind; observe] the scouting plane put in five more appearances, and to feel the certainty of this dread rising within him.	

Table 4: Example predictions based on nearest neighbor sentences. The word in question is marked in boldface, subset with a short description of its WordNet synset (true positives green, false positives red).

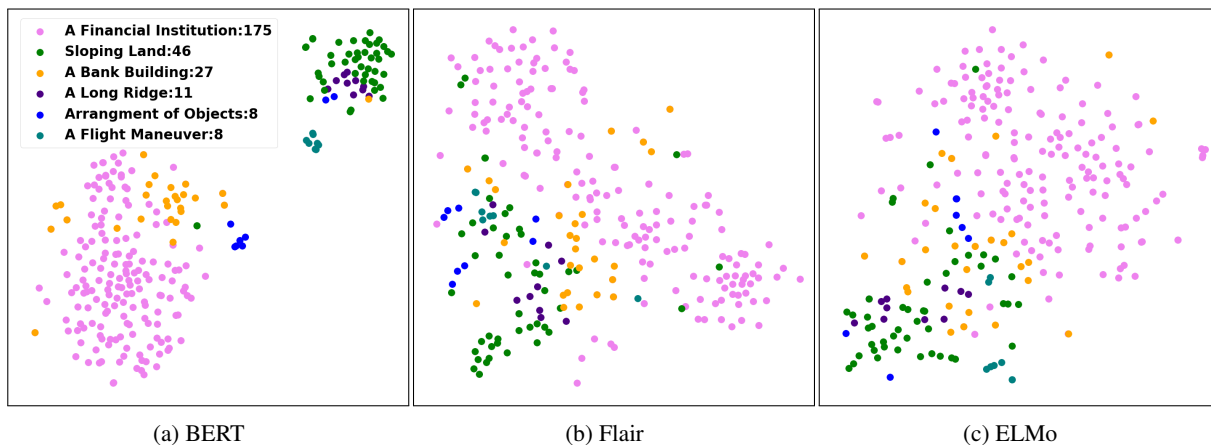


Figure 1: T-SNE plots of different senses of ‘bank’ and their contextualized embeddings. The legend shows a short description of the respective WordNet sense and the frequency of occurrence in the training data. Here, the SE-3 training dataset is used.

k	SE-2	SE-3	S7-T7		S7-T17	
			SemCor	WNGT	SemCor	WNGT
1	76.10	78.62	79.30	85.23	61.38	61.98
3	76.52	79.66	79.44	85.01	60.94	62.64
7	76.50	80.02	79.35	85.05	62.50	62.20
10	76.40	80.12	79.40	85.10	62.72	62.20
30	76.43	79.66	79.40	85.14	63.17	61.98
70	76.43	79.61	79.35	85.23	62.95	61.98
100	76.43	79.63	79.35	85.32	62.95	61.98
300	76.43	79.63	79.35	85.32	62.95	61.98

Table 5: Best POS-sensitive kNN models. Bold numbers are improved results over Table 3.

representation into different regions of the shared vector space, we used a k-nearest neighbor approach, which allows us to investigate their properties on an example basis. For experimentation, we used pre-trained models from Flair NLP (Akbi et al., 2018), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2019). Further, we tested our hypothesis on four standard benchmark datasets for word sense disambiguation. We conclude that WSD can be surprisingly effective using solely CWEs. We are even able to report improvements over state-of-the-art results for the two lexical sample tasks of SenseEval-2 and SensEval-3.

Further, experiments showed that CWEs in general are able to capture senses, i.e. words, when used in a different sense, are placed in different regions. This effect appeared strongest using the BERT pre-trained model, where example instances even form clusters. This might give rise to future directions of investigation, e.g. unsupervised word sense-induction using clustering techniques.

	Noun	Adj	Verb	Other	avg #POS per word
SE-2 (tr)	41.32	16.57	42.11	0.00	1.0
SE-2 (te)	40.98	16.56	42.46	0.00	1.0
SE-3 (tr)	46.45	3.94	49.61	0.00	1.0
SE-3 (te)	46.17	3.98	49.86	0.00	1.0
S7-T7	49.00	15.75	26.14	9.11	1.03
S7-T17	34.95	0.00	65.05	0.00	1.01
SemCor	38.16	14.70	38.80	8.34	1.10
WNGT	57.93	21.57	15.55	4.96	1.07

Table 6: Percentage of senses with a certain POS tag in the corpora.

Since the publication of the BERT model, a number of extensions based on transformer architectures and language model pre-training have been released. In future work, we plan to evaluate also XLM (Lample and Conneau, 2019), RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019) with our approach. In our qualitative error analysis, we observed many near-misses, i.e. the target sense and the predicted sense are not particularly far away. We will investigate if more powerful classification algorithms for WSD based on contextualized embeddings are able to solve this issue even in cases of extremely sparse training data.

Acknowledgments

This work was funded by BWFG Hamburg in the Forum 4.0 project, by DFG in the JOIN-T 2 project and by DAAD in form of a WISE stipend.

References

- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, NM, USA.
- Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 130–138, Cadiz, Spain.
- Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63(1):743–788.
- Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, MN, USA.
- 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.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.
- Zellig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.
- Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal.
- Ganesh Jawahar, Benoît Sagot, and Djamel Seddah. 2019. What does BERT learn about the structure of language? In *57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy.
- Mikael Kågebäck and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional LSTM. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 51–56, Osaka, Japan.
- Adam Kilgarriff. 2001. English lexical sample task description. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17–20, Toulouse, France.
- Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany.
- Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013*, Scottsdale, AZ, USA.
- George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.
- George A. Miller, Claudia Leacock, Randee Teng, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology, HLT '93*, pages 303–308, Princeton, NJ, USA.
- Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval '07*, pages 30–35, Prague, Czech Republic.
- Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar.
- Alexander Panchenko, Fide Marten, Eugen Ruppert, Stefano Faralli, Dmitry Ustalov, Simone Paolo Ponzetto, and Chris Biemann. 2017. Unsupervised, knowledge-free, and interpretable word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 91–96, Copenhagen, Denmark.

- Michel Pêcheux, Tony Hak, and Niels Helsloot, editors. 1995. *Automatic discourse analysis*. Rodopi, Amsterdam and Atlanta.
- Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183, Berlin, Germany.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, New Orleans, LA, USA.
- Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval '07*, pages 87–92, Prague, Czech Republic.
- Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. 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.
- Steffen Remus and Chris Biemann. 2018. Retrofitting Word Representations for Unsupervised Sense Aware Word Similarities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.
- Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1793–1803, Beijing, China.
- Ferdinand de Saussure. 2001. *Grundfragen der allgemeinen Sprachwissenschaft*. de Gruyter, Berlin, 3 edition.
- 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.
- Kaveh Taghipour and Hwee Tou Ng. 2015. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–323, Denver, CO, USA.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.
- Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018a. Improving the coverage and the generalization ability of neural word sense disambiguation through hypernymy and hyponymy relationships. *CoRR*, abs/1811.00960.
- Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018b. UFSAC: Unification of Sense Annotated Corpora and Tools. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.
- Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. *CoRR*, abs/1905.05677.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.
- Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1374–1385, Osaka, Japan.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffenseEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, MN, USA.