Diachronic Sense Modelling: Tracking of sense gains and losses over time

Tony Nguyen

Abstract

This paper builds upon previous works in using word embeddings to study how words change over time. Leveraging the effectiveness of novel state-of-the-art neural language architectures stemming from the Transformer (Vaswani et al., 2017), the paper examines a framework of analyzing word sense frequencies using sense embeddings to detect gains and losses of senses of words over time, as well as extend previous methods of using diachronic word embeddings to automatically detect semantic change of the English lexicon.

1 Introduction

The field of natural language processing has evolved tremendously over the years, especially with the emergence of many neural architectures. From the development of recurrent neural networks (RNN) (Williams et al., 1986) aiming to capture the long-term dependencies in language modelling to the long-short term memory (LSTM) (Hochreiter et al., 1997) architecture that addresses RNN's vanishing gradient problem, state-of-the-art results continue to beat one another. Recent advancements in this field have given rise to the novel Transformer (Vaswani et al., 2017). This language architecture uses scaled-dot product attention to learn dependencies between tokens of the same sentence (self-attention), as well as dependencies between the decoded output and the given input (encoder-decoder attention). By performing this attention mechanism in a bi-directional manner, models using this architecture are able to be more robust at understanding relationships between words, as each word is able to "see" its entire context. In comparison, RNNs and LSTMs sequence-to-sequence architectures have both autoregressive encoders and decoders, hence each token only sees previous tokens as context. Using the Transformer, state-of-the-art results have been produced in many downstream language processing tasks such as machine translation or text generation. Since then, many pre-training techniques had been proposed, the pioneering one most probably being BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) (Devlin et al., 2018) which uses a stack of Transformer encoders to train sentences on the masked language modelling objective.

BERT has proven to be successful at performing many tasks such as summarization or question answering. Recently, it has been applied to the field of semantic change where we study the shift in meaning of words as time progresses. Specifically, words can be polysemous or have different senses at different instances. Depending on the cultural influence at different time periods, a particular sense of a word may be used with varying frequencies if we examine its usage through history. These frequencies can be obtained by counting the number of times each sense of a word is used after the correct senses have been disambiguated using sense inventories. In natural language processing, this task is known as "word sense disambiguation" (WSD). In the aforementioned work (Hu et al., 2019), the authors propose the use of "sense embeddings" from BERT as an approach to this task of sense modelling, citing the effectiveness of the deep contextualized word embeddings which it produces. As mentioned previously, BERT weighs the static word embeddings using the self-attention mechanism, hence the output encoded representation of a word in a particular sentence is contextualized with all tokens from the sentence.

This paper proposes and explores a pipeline which

builds upon the use of sense embeddings for tracking the diachronic change in frequencies of sense usage. From this, we deduct whether or not a particular sense of a word has been lost or gained. In addition, embeddings at different time periods for the same word (diachronic word embeddings) have been extracted and compared to deduce a degree of semantic change. These frameworks for automatic semantic change detection utilize older methods of word embeddings extraction including local mutual information (LMI) scores co-occurrence matrix (Gulordava et al., 2011), positive pointwise mutual information (PPMI) co-occurence matrix and skip-gram with negative sampling (perhaps more commonly known as word2vec) (Hamilton et al., 2016), (Mikolov et al., 2013). A natural extension of this approach is to use the higher quality word embeddings from BERT to measure lexical semantic change degrees, which this paper will also examine.

2 Related Work

2.1 Word Sense Disambiguation

The work in this paper builds upon work done previously in different sub-fields of natural language processing. The task of word sense disambiguation has been studied extensively ever since its introduction. The problem is difficult, being "AI-complete" (Navigli, 2009), (Ide et al., 1998). Previous approaches to word sense disambiguation include supervised, semi-supervised, unsupervised, and knowledge-based methods (Navigli, 2009). Supervised approaches include decision trees using the ID3 algorithm, Naïve Bayes classifiers and neural networks. With supervised methods, however, there is a bottleneck that is the lack of labelled data for the task of sense disambiguation (Navigli, 2009). Semi-supervised methods aim to alleviate this limiting factor by using a small labeled dataset to create a larger dataset using techniques such as bootstrapping (Hearst, 1991). A recent semi-supervised method is to apply an LSTM on labeled data, and then use label propagation to annotate unlabeled data by computing similarity metrics between the unlabeled and labeled data (Yuan et al., 2016). Unsupervised methods use the key idea that instances of a word that have the same sense have similar neighbors (Navigli, 2009). Knowledge-based approaches include the incorporation of our semantic knowledge into computational methods of disambiguating senses. For instance, disambiguating a sense of a word in a sentence heavily relies on the usage of the surrounding context. However, not all context tokens are equally important and our algorithms should focus on thematic words instead of syntactic words (O et al., 2018). It is noteworthy that the study of WSD algorithms has been made possible by resources such as WordNet (Fellbaum, 1998).

2.2 Diachronic Sense Modelling

Extensive work has been done in the investigation of how words change over time. However, most of these approaches consider words as whole (Gulordava et al., 2011) and disregard the polysemous nature of some words. That is, words can have different senses when used in different contexts. Attempts at detecting word sense changes include usage of novelty measurements in detecting novel senses, computed by comparing a reference corpus with a new corpus (Lau et al., 2012). Other approaches in diachronic sense modelling include co-occurrence graph clustering (Mitra et al., 2014) and curvature clustering (Tahmasebi et al., 2017). The former method uses the Chinese Whispers algorithm to produce a set of clusters for each target word by decomposing its neighboring words. The authors hypothesized that these clusters represent the senses of the word and by comparing the clusters of the same words between time periods, can be detected. More recently, the application of BERT to this topic was proposed (Hu et al., 2019). The authors used a pre-trained BERT model to extract "sense embeddings" by averaging out word embeddings of all instances of a word with the same sense. Frequency scores are then calculated using corpuses from varying time periods to study the frequency variation of senses over time. In this paper, this method of using sense embeddings to measure sense use frequencies will be applied to the tracking of sense gains and losses over time.

3 Methodology

This section outlines and elaborates upon a pipeline to detect sense gains and losses, as well as two approaches to measuring degree of semantic drift of words.

3.1 Tracking of sense gains and losses

The fundamental idea of the approach is to obtain measurements of frequencies of a sense of a word

at different time periods. The frequency of a sense at a time period indicates how often that sense is used. By comparing these values across different time periods, changes to sense usages can be observed.

To measure the frequency of a sense of a polysemous word, many instances of the word need to be collected. For each of these instances, the correct sense needs to be disambiguated (using a WSD approach). This pipeline uses the method of comparing sense embeddings with the target word's embedding to identify the sense of an instance of a word (Hu et al., 2019). To obtain these sense embeddings, a pre-trained BERT model is finetuned on the WSD task. After disambiguating senses for all collected instances of a word, the frequencies can be calculated as a normalized occurrence score. By doing this across data obtained from different time periods, the frequency of the same sense of a word at different times can be examined for any change. A threshold value is introduced to determine whether or not a sense is "present". By comparing the presence of senses of words across different times, sense gains and losses can be detected.

More specifically, this pipeline follows the following steps in order:

1. **Finetune a pre-trained BERT model on the WSD task**: Due to computational limitations, the full BERT model could not be used. The results in this paper are obtained from the DistilBERT architecture implemented and pre-trained by Hugging Face Co., which is much smaller in size while retaining 97% of BERT's original performance (Sanh et al., 2019). The model receives as input a combined sentence and a sense from the sense inventory of the target word and performs a binary classification, indicating whether or not the target word in the given sentence has the given sense. Figure 1 illustrates the finetuning architecture.

   The DistilBERT final hidden outputs are ran through 2 affine layers with Gaussian Error Linear Unit (GELU) non-linearity (Hendrycks et al., 2016) before projected to the labels space. The fully connected layers' weights are initialized using Xavier initialization (Glorot et al., 2010) and the biases are initialized to 0. Due to computational limitations, the only parameters being finetuned are those in the classifier and the embedding weights in DistilBERT. Ideally, the final encoder layer should also be finetuned.

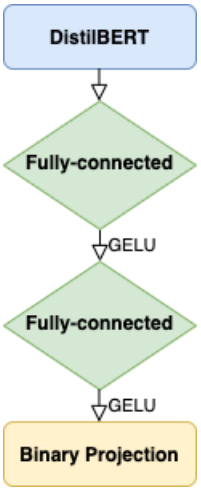

Figure 1: DistilBERT WSD finetune architecture

2. **Collecting sense embeddings**: For a sense s_i of a word w_i, its sense embedding $e^{w_i s_i}$ can be computed to be the average of the word embeddings of w_i collected from $\{S_1, S_2, ..., S_n\}$ where each S_k is a sentence containing the word w_i being used with the sense s_i. These sentences are taken from a corpus where sentences with target words are labelled with the correct sense. Due to the nature of Transformer-based architectures like BERT, the word embedding of the same word w_i will be attention weighed differently given different context. The word embedding of a word w_i in a sentence S_k can be extracted by running the sentence through the finetuned DistilBERT model and accessing the final hidden outputs (Hu et al., 2019).

3. **Identifying senses and computing frequencies**: After sense embeddings have been obtained, a different corpus that is unlabelled is used to measure sense use frequencies. For each target word of interest, all sentences containing the target word are collected. To get the sense of w_i in a sentence S_k, cosine similarity scores are computed between the w_i's embedding in S_k and all of w_i's sense embeddings. The sense s_k of the word w_i

in sentence S_k is the one wtih the maximal aforementioned cosine similarity score. In other words:

$$s_k = \underset{s_i}{\text{argmax}} \frac{e^{w_i s_i} \cdot e_k^{w_i}}{\|e^{w_i s_i}\|_2 \|e_k^{w_i}\|_2}$$

where $e^{w_i s_i}$, as defined in the previous step, is the sense embedding of the sense s_i of word w_i and $e_k^{w_i}$ is the word embedding of the word w_i in the sentence S_k (Hu et al., 2019). After identifying the correct sense for all sentences containing a target word, the frequency scores for the senses can be measured.

A sense s_i of a word w_i has the frequency score $\mathcal{F}_{w_i s_i}$ which can be computed using:

$$\mathcal{F}_{w_i s_i} = \frac{\mathcal{N}_{w_i s_i}}{\sum_k \mathcal{N}_{w_i s_k}}$$

where $\mathcal{N}_{w_i s_i}$ is the number of times the word w_i is used with the sense s_i (Hu et al., 2019). Essentially, the frequency score of a sense of a word is the ratio of the number of times that particular sense is used and the total number of times the word is used.

4. **Identifying sense gains and losses**: As mentioned above, a threshold, in this case 2%, is used to identify if a sense is a present or not. To identify sense gains and losses, step 3 is performed on 2 corpuses from 2 different time periods to obtain frequencies of all senses of all words across the time periods. If the frequency of a sense of a word at a time period falls below this threshold value, the sense is considered to be insignificant and the corresponding word is considered to not have the sense. Otherwise, the sense is considered present and the corresponding word is considered to have that sense at that time period. If a sense of a word is considered to be not present at a previous time period, but becomes present in a later time period, the word is said to have gained that sense. Similarly, a word is said to have lost a sense if the sense frequency falls below the threshold at a later time after being at least the threshold at a previous time.

3.2 Measuring degree of semantic drift

In this paper, two methods for detecting semantic change will be discussed. The first measures a divergence score between the frequencies of senses of words between time periods and the second extends upon previous works that used diachronic word embeddings.

3.2.1 Sense frequencies divergence

To measure the degree of lexical semantic change, the Jensen-Shannon divergence (JSD) between the frequencies of a word between time periods was computed. Because these frequency scores, obtained as outlined in section **3.1**, are normalized occurrence counts, they can be interpreted as probability distributions. Thus, the divergence between two probability distributions P and Q can be computed using $JSD(P, Q)$. JSD is preferred over Kullback-Leibler divergence (D_{KL}) due to the symmetry the former has which the latter lacks. In other words, $JSD(P, Q) = JSD(Q, P)$ while $D_{KL}(P \mid\mid Q) \neq D_{KL}(Q \mid\mid P)$ (Schlechtweg et al., 2020).

3.2.2 Diachronic word embeddings

Another method to measure semantic drift is to utilize word embeddings obtained at different time periods and compute a distance metric between these embeddings. With architectures like BERT yielding state-of-the-art results, the word embeddings from the embedding layer of these architectures are believed to be of higher quality, evident from better results when using them on downstream tasks such as text classification. This framework of measuring semantic change takes advantage of these higher quality embeddings to extend upon previous works that measured semantic change by using diachronic word embeddings. The following steps are taken:

1. 2 pretrained DistilBERT models are fine-tuned on the masked language modelling task using 2 corpuses from 2 different time periods using implementations from Hugging Face Co. (Sanh et al., 2019).

2. Embeddings of the same word are extracted from the embedding layer of the DistilBERT model. These embeddings are 768-dimensional vectors. Using these embeddings as is did not yield reasonable results, most probably due to the curse of dimensionality. To alleviate this problem, the em-

beddings' dimensionality are reduced into 2-dimensional space using the t-distributed stochastic neighboring embedding (t-SNE) algorithm (Maaten et al., 2008).

3. These dimensionality-reduced vectors can be used to compute cosine distance scores which represent the corresponding words' degree of semantic drift.

4 Datasets

4.1 Finetuning BERT for WSD

To finetune the pre-trained BERT model from Hugging Face Co., the SemCor corpus was used. This specific version was formatted to match the format used in SemEval competitions (Raganato et al., 2017). This dataset contains sentences with target words tagged with their WordNet 3.0 senses and part-of-speech (POS). WordNet is a lexical database containing word senses built by Princeton (Fellbaum, 1998). All words in all sentences are given in their lemmatized form. From these sentences, a JSON file that maps a training sentence to all ambigious target words within it and the associated POS was compiled. For each target word in the dataset, all of its senses are scraped from WordNet and a dictionary of words to senses were constructed. From this word to senses dictionary as well and the JSON file, a dataset of approximately 2 million examples was built, combining a sentence with a target word's sense for binary classification. The true sense was provided for a target word of a sentence, and thus the constructed dataset labels this combination of word, sense and sentence as correct (1) and all other senses with the same word and sentence are labelled as incorrect (0). It is noteworthy that this setup leads to the dataset containing more incorrect examples than correct ones because for a combination of word and sentence, there is only one correct sense, but possibly multiple incorrect senses.

4.2 Sense embeddings collection

To collect sense embeddings, sentences with words having particular senses must be obtained. For this task, the combined SemCor and OMSTI dataset was used (Raganato et al., 2017). This combined dataset contains the same examples as the the SemCor dataset described in part **4.1**, but with extra examples from the OMSTI dataset. In the interest of using the evaluation framework available, which analyzes word sense change for a particular POS of a word, only sentences of target words available in the evaluation set with the appropriate sense POS are considered. For example, the evaluation dataset analyzes the sense change of the word "attack" as a noun. Hence, all sentences containing the word "attack" being used as a noun are collected while those using the word "attack" as a different POS are omitted.

4.3 Sense frequencies analysis and diachronic word embeddings extraction

These two tasks use the same dataset as they merely apply previous computations to time varying datasets. The dataset used in these two tasks are cleaned, formatted, preprocessed and lemmatized versions of the COHA corpus from two different time periods (Alatrash et al., 2020). The first corpus has text collected from the time period 1810 - 1860 while the second has text collect from 1960 - 2010. For the task of sense frequencies analysis, only sentences containing target words present in the evaluation dataset are used for evaluation purposes. For the task of extracting diachronic word embeddings, the entire text of both corpora is used.

4.4 Evaluation data

Gold standard data from the SemEval2020 competition was used to evaluate this framework. To evaluate the pipeline for tracking sense gains and losses, a binary word sense change dataset was used. The dataset contains humans-annotated labels of whether or not a word has "changed" between the two corpora. If a word has lost or gained a sense between the time periods, the label will be 1 (indicating that the word's sense(s) has(have) changed). If none of a word's sense has been lost and the word has not gained a new sense, the word's label will be 0.

The framework's semantic change degree are compared against human annotated degrees of change using a Pearson correlation coefficient. The resulting semantic change ranking is also compared against the human annotated ranking of semantic change of target words using a Spearman's rank-order correlation coefficient (Schlechtweg et al., 2019).

WordNet 3.0 sense	t_1 frequency	t_2 frequency
the act of attacking	0.0022	0.0060
(military) an offensive against an enemy (using weapons)	0.1806	0.4202
ideas or actions intended to deal with a problem or situation	0.0220	0.0060
intense adverse criticism	0.6476	0.3169
a decisive manner of beginning a musical tone or phrase	0.0859	0.0756
a sudden occurrence of an uncontrollable condition	0.0595	0.1741

Table 1: Sense frequencies for the word "attack" as a noun at time periods t_1 (1810 - 1860) and t_2 (1960 - 2010). The first column are all WordNet 3.0 senses of the word "attack" that are nouns. The second and third columns are frequencies represented by the normalized occurrence counts of each sense in t_1 and t_2, respectively. These frequencies are obtained as outlined in section **3.1**.

5 Results

5.1 Word sense disambiguation

A holdout set from the dataset described in **4.1** was used to evaluate the performance of BERT on the WSD task. 1000 batches of WSD examples from this holdout set was used, where each batch containing 64 examples. An accuracy score of 91.4% was achieved. Although this score is impressive, it is noteworthy that the way the dataset was built for WSD finetuning, the task simply becomes binary classification with a lot of false examples. Therefore, although the accuracy is high, it does not indicate anything out of the ordinary.

Two examples of successful sense disambiguation for the word "ball" are shown below. The text following the colon is the WordNet 3.0 sense of the target word disambiguated by the model.

1. The player kicked the **ball** way too hard: "round object that is hit or thrown or kicked in a game"

2. Harry went to the Yule **Ball** in his fourth year: "the people assembled at a lavish formal dance"

5.2 Diachronic sense modelling

Table 1 shows the sense frequencies for the word "attack" as a noun during two different time periods: 1810 - 1860 (t_1) and 1960 - 2010 (t_2). From the table, it is observable that during the period t_1, the dominant sense of the word "attack" is "intense adverse criticism" as the word "attack" is used with this sense 64.76% of the time. However, in time period t_2, the dominant sense is shifted to "(military) an offensive against an enemy (using weapons)". Using the threshold of 2% as discussed previously, the sense "ideas or actions intended to deal with a problem or situation" is considered to be lost through the time periods as in t_1, the sense was used with frequency 2.2% while in t_2, this value dropped to 0.60%. The word did not gain any sense by this metric. The sense "the act of attacking" would be considered to not be present at either time periods since its frequency was below 2% in both t_1 and t_2.

To numerically evaluate this approach for detecting sense gains and losses, each word was classified as either changed (1) or unchanged (0). A word is classified as changed if and only if it has either gained or lost any sense. The classification accuracy is reported below with other systems' scores. Due to the approach used and the lack of data, some words from the evaluation set were omitted. Therefore, two scores are being reported below. SE is the score computed using only words available both in the evaluation set and the computations. SE-penalized considers all omitted words as being incorrect.

System	Accuracy
NLPCR	73.0%
RPI-Trust	70.3%
SE	69.0%
Baseline	59.5%
SE-penalized	54.0%

Table 2: Word change classification accuracies

5.3 Lexical semantic change

Figure 2 shows two graphs plotting word embeddings of words that are classically used to study

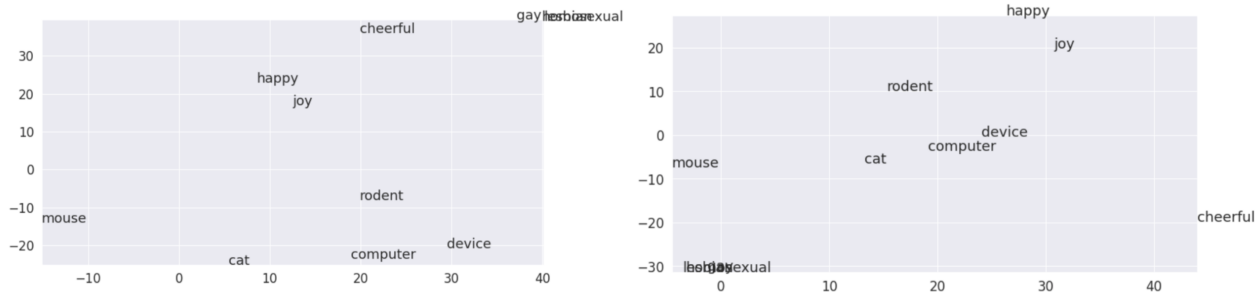

Figure 2: Word embeddings obtained from DistilBERT finetuned on corpuses of different time periods

semantic change. The word embeddings obtained from the DistilBERT models are 768-dimensional and hence cannot be plot in 2-dimensional space. The plotted coordinates are vectors obtained after running t-SNE on the 768-dimensional vectors to reduce their dimensionality to 2 components. From examining the graph on the left, the word "gay" can be seen to have very close semantic meaning to the words "homosexual" and "lesbian", but also words "cheerful", "happy" and "joy". The word "mouse" and "cat" are relatively close in semantic distance. In comparison, in the graph on the right, the words "gay", "lesbian" and "homosexual" are still very close in meaning. However, the word "gay" is no longer close in meaning to "happy", "joy" or "cheerful" as it used to in the previous time period. Furthermore, the word "mouse" has moved closer in meaning to "computer" and "device".

The approaches explored for computing degree of lexical semantic change are evaluated by computing a Pearson correlation coefficient (PCC) between the actual degrees of change and the expected degrees of change. Furthermore, the actual and expected ranking of the degrees of change are used to compute a Spearman-rank correlation coefficient (SPR). The results are shown in the following table.

System	PCC	SPR
DWE + t-SNE	0.405	0.463
NLPCR	-	0.440
UG Student Intern	-	0.422
SE + JSD	0.467	0.266

Table 3: Correlation coefficients between expected and actual degrees of lexical semantic change

DWE + t-SNE is the result of the method using diachronic word embeddings reduced to 2-dimensional space using t-SNE. SE + JSD is the result from using sense embeddings to collect frequencies and measure the Jensen-Shannon divergence between the frequencies of senses of words between the time periods.

6 Discussion

6.1 Sense gains and losses

From the previous section, most of the times the word "attack" was used between 1810 - 1860, it was used the sense of an "intense adverse criticism". In comparison, between 1960 - 2010, the word was mostly used with the sense of a military attack using weapons. This shift in the dominant sense can be most probably be attributed to the World Wars which occurred in the early 1900s. Furthermore, the sense "ideas or actions intended to deal with a problem or situation" can be seen to be lost between the time periods. This sense loss may be due to the availability of more common words to express the same meaning such as "approach" or "method".

The pipeline explored in this paper yielded decent results for the tracking of sense gains and losses. However, the evaluation method was limited as the test set comes from human annotated word sense change. The metrics that humans use can be subjective and thus the numeric values proposed to be tested against can be inaccurate and misleading. Hence, a better evaluation method should be investigated. One such method is to use historical dictionaries such as the Concise Oxford English Dictionary of different editions which document contemporary usage of words (Lau et al., 2012). Definitions of a word can be viewed as its senses. Therefore, by comparing these defini-

tions across different editions published at different times, losses and gains of a particular sense can be detected by checking for the presence of a previously absent sense, and vice versa.

6.2 Lexical semantic change

From the graphs presented in Figure 2 which plot the word embeddings obtained at different time periods, the general expected behavior of semantic shifted can be observed. The word "gay" does hold close meaning to words "lesbian" and "homosexual". However, previously, it also means "joy" or "happy", though that meaning of the word "gay" has been lost during recent times. This semantic shift of the word "gay" is observed in the graphs in Figure 2. Furthermore, the word "mouse" is commonly used to describe the small rodent. With the growing prevalence of technologies, the word has adopted a new meaning to describe the device that is used to control laptops, computers, etc. The adoption of this new meaning by the word "mouse" can be seen in the graph as its vector representation moves closer to that of "device" and "computer".

The evaluated results of the usage of BERT's embeddings to compute degrees of lexical semantic change seem promising. However, this result is obtained after running the t-SNE dimensionality reduction algorithm multiple times. Because the algorithm initializes the learned distribution randomly, random seeding with different values can yield different results. When using the sense embeddings technique to compute frequencies of senses at different time periods and measuring JSD between these frequencies, the sizable discrepancy between the two correlation coefficients are observed, which can be a result of the presence of outliers. Hence, the Pearson correlation coefficient may not be representative of the association between the predicted semantic change degrees and the expected ones. Hence, it can be concluded that the ranking of the words' semantic change degree computed using this approach does not correlate very strongly with that decided by judges. However as discussed in section **6.1**, human judges' scores might not be entirely accurate.

6.3 Conclusion and Future Works

Words can be polysemous, thus having different senses when used in different contexts. A word as a whole entity also change semantically over time. Through this investigation, it is evident that BERT can be applied to detect how words change over time, as the proposed framework is able to detect word senses gains and losses as well as lexical semantic change. By computing sense embeddings of senses of words and using them to disambiguate a word sense in a given context, frequency scores can be obtained for word senses. These frequencies can be compared across time periods to detect change in sense usage, especially gains and losses. In addition, distance measurements between word embeddings from BERT obtained at different time periods can be used to measure the degree of semantic change.

Drawbacks of BERT that can influence its performance here include the discrepancy between pretraining and finetuning task. In other words, in this pipeline of detecting sense gains and losses, BERT was finetuned on the WSD task which is very different from the masked language modelling task which the model was pre-trained on. To train on the masked language modelling objective, the model corrupts incoming data and attempts to corrupted tokens. In the task of WSD or any other real-life natural language processing tasks, corrupted data is typically not available. A possible future investigation is to use an architecture such as XLNet which solves this problem of BERT (Yang et al., 2019).

References

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

[Williams et al.1986] Ronald J. Williams, Geoffrey E. Hinton and David E. Rumelhart. 1986. Learning representations by back-propagating errors. Nature. 323 (6088): 533–536.

[Hochreiter et al.1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short Term Memory. *Neural computation*, 9(8):1735–1780, 1997.

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

[Hu et al.2019] Renfen Hu, Shen Li and Shichen Liang. 2019. Diachronic Sense Modeling with Deep Con-

textualized Word Embeddings: An Ecological View. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899-3908.

[Gulordava et al.2011] William L. Hamilton, Jure Leskovec and Dan Jurafsky. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, pages 67-71.

[Hamilton et al.2016] William L. Hamilton, Jure Leskovec and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489-1501.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

[Navigli2009] Roberto Navigli. 2009. Word sense disambiguation: A survey. In *ACM Computing Surveys*, Vol. 41, No. 2, Article 10.

[Ide et al.1998] Nancy Ide and Jean Véronis. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. In *Computational Linguistics*, Vol. 24, No. 1, pages 1-40.

[Hearst1991] Marti A. Hearst. 1991. Noun Homograph Disambiguation Using Local Context in Large Text Corpora. In *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, Oxford.

[O et al.2018] Dongsuk O, Sunjae Kwon, Kyungsun Kim and Youngjoong Ko 2018. Word Sense Disambiguation Based on Word Similarity Calculation Using Word Vector Representation from a Knowledge-based Graph. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2704-2714.

[Yuan et al.2016] Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans and Eric Altendorf. 2016. Semi-supervised Word Sense Disambiguation with Neural Models. In *CoRR*.

[Fellbaum1998] Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[Lau et al.2012] Jey Han Lau, Paul Cook, Diana McCarthy, David Newman and Timothy Baldwin. 2012. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591-601.

[Mitra et al.2014] Sunny Mitra, Ritwik Mitra, Martin Ried,Chris Biemann, Animesh Mukherjee and Pawan Goyal. 2014. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pages 1020-1029.

[Tahmasebi et al.2017] Nina Tahmasebi and Thomas Risse. 2017. Finding Individual Word Sense Changes and their Delay in Appearance. In *RANLP*, pages 741-749.

[Sanh et al.2019] Victor Sanh, Lysandre Debut, Julien Chaumond and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[Hendrycks et al.2016] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian Error Linear Units (GELUs). In *CoRR*.

[Glorot et al.2010] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[Schlechtweg et al.2020] Dominik Schlechtweg and Sabine Schulte im Walde. 2020. Simulating Lexical Semantic Change from Sense-Annotated Data. In *The Evolution of Language: Proceedings of the 13th International Conference (EVOLANGXIII)*.

[Maaten et al.2008] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. In *Journal of Machine Learning Research 1*, pages 1-48.

[Raganato et al.2017] Alessandro Raganato, Jose Camacho-Collados and Roberto Navigli. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of EACL 2017, Valencia, Spain*.

[Alatrash et al.2020] Reem Alatrash, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020. Clean Corpus of Historical American English (CCOHA). In *Proceedings of LREC*.

[Schlechtweg et al.2019] Dominik Schlechtweg, Anna Hätty, Marco del Tredici and Sabine Schulte im Walde. 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[Yang et al.2019] 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*.