Non-Parametric Word Sense Disambiguation for Historical Languages

Enrique Manjavacas
Leiden University
Leiden, The Netherlands
enrique.manjavacas@gmail.com

Lauren Fonteyn
Leiden University
Leiden, The Netherlands
l.fonteyn@hum.leidenuniv.nl

Abstract

Recent approaches to Word Sense Disambiguation (WSD) have profited from the enhanced contextualized word representations coming from contemporary Large Language Models (LLMs). This advancement is accompanied by a renewed interest in WSD applications in Humanities research, where the lack of suitable, specific WSD-annotated resources is a hurdle in developing ad-hoc WSD systems. Because they can exploit sentential context, LLMs are particularly suited for disambiguation tasks. Still, the application of LLMs is often limited to linear classifiers trained on top of the LLM architecture. In this paper, we follow recent developments in non-parametric learning and show how LLMs can be efficiently fine-tuned to achieve strong few-shot performance on WSD for historical languages (English and Dutch, date range: 1450-1950). We test our hypothesis using (i) a large, general evaluation set taken from large lexical databases, and (ii) a small real-world scenario involving an ad-hoc WSD task. Moreover, this paper marks the release of `GysBERT`, a LLM for historical Dutch.

1 Introduction & Related Work

A common task in Natural Language Processing (NLP) applications is the disambiguation of a particular target word in a given context. Due to a variety of reasons (see e.g. Blank, 1999), word forms may be semantically extended to a range of different meanings or word senses (e.g. *rat* 'animal' > 'informer, snitch'). Automated disambiguation—i.e. the mapping of an ambiguous word form to its intended underlying word sense—is a task that can help in many different types of information extraction and text mining tasks.

In recent years, WSD approaches have shifted towards contextualized embeddings extracted from Large Language Models (LLM) like BERT (Devlin et al., 2019). These token-based embeddings incorporate substantial semantic information from the target lexical item and the sentential context that surrounds it. For this reason, LLMs are particularly well suited to disambiguation tasks and have, in fact, already been shown to indirectly capture word senses and perform competitively (Reif et al., 2019; Hadiwinoto et al., 2019). More recently, LLMs also obtained state-of-the-art performance using semantic networks (Loureiro and Jorge, 2019) and gloss information (Luo et al., 2018; Huang et al., 2019; Blevins and Zettlemoyer, 2020).

These recent advancements in WSD by means of LLMs are also interesting for Humanities scholars, as large-scale corpus-based studies have steadily become more standard practice over the last decades. Traditionally, data annotation in Humanities research is done manually, but with ever-growing size of corpora, such manual WSD has become decreasingly feasible. At the same time, researchers are also increasingly interested in disambiguating word senses without relying on intuitive judgment, which is ultimately subjective. This desire for a more data-driven approach to WSD is particularly prominent in research involving historical language, for which researchers are unable to solicit native speaker interpretations and expert annotators are rare. There is, in short, a growing interest in and need for automated WSD in Humanities research, and researchers have turned to NLP techniques to meet their needs.

As examples, we find projects aiming to trace the history of concepts over time (e.g. Beelen et al., 2021), automatically detect (different types of) lexical-semantic (e.g. Sagi et al., 2011; Giulianelli et al., 2020) and grammatical change (e.g. Fonteyn, 2020) or to quantify the evolution of the senses of a word (e.g. Tahmasebi et al., 2018). Because of their temporal dimension, such projects require NLP architectures that can process word senses in settings which are complicated by an evolving grammar and lexicon (Fonteyn, 2020), shifting spelling conventions, and noise introduced by

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 123–134
November 20, 2022.

OCR errors (Piotrowski, 2012). Yet, applications of LLMs on historical corpora tend to be either limited to arithmetic operations on the vector space of frozen contextual embeddings, or restricted to fine-tuning a linear layer to perform disambiguation to a pre-determined number of senses (Hagen et al., 2020; Beelen et al., 2021; Manjavacas and Fonteyn, 2021).

In this paper, we explore the capabilities of Large Language Models (LLMs) á la BERT (Devlin et al., 2019) for WSD on historical text. In doing so, we investigate to what extent (i) fine-tuning can improve WSD over arithmetic operations on plain frozen embeddings, and (ii) how much annotated data is needed in order to obtain gains over a baseline. Our focus lies on examining the data efficiency of LLM-based approaches for WSD, because annotated resources for WSD are generally scarce and costly to generate – a problem that is exacerbated with historical languages, where rare expert historical knowledge is required to produce annotated resources. To attain these goals, we take inspiration from recent metric-based non-parametric approaches (Holla et al., 2020; Du et al., 2021; Chen et al., 2021). These aim to optimize a model on a set of learning tasks (e.g. the disambiguation of a given ambiguous word) so that the model can quickly adapt to perform well on similar future tasks (e.g. disambiguating sentences of an ambiguous new word on the basis of a small annotated set of word senses). More specifically, we deploy a non-parametric approach to WSD fine-tuning that does not rely on additional task-specific parameters and that achieves surprisingly strong performance on out-of-domain lemmas. We argue that these results are promising if we aim to extend the scope of applications of WSD models in the Humanities.

Main Contributions Our experiments show that a metric-based parameter-free approach to few-shot WSD can achieve promising performance on historical data, even on held-out lemmas that were not seen during training. To do so, they require only a small number of training lemmas and word sense examples.

Moreover, we show that historical pre-training can push performance even further. To this end, we rely on `MacBERTh` (Manjavacas and Fonteyn, 2021, 2022), a LLM pre-trained on historical English. Additionally, in order to back up the results across different languages, a new historically pre-trained LLM for Dutch named `GysBERT` was developed and tested. The release of `GysBERT` accompanies the publication of the present study.[1]

Outline In Section 2, we describe the architecture used in order to tackle WSD in a non-parametric way. Subsequently, in Section 3, we describe the resources, datasets and pre-trained models underlying the present study. In Section 4, we present a series of experiments in order to illustrate the main results achieved by the evaluated approaches, focusing on small training regimes in Section 4.2, as well as the effect of time in Section 4.3. Section 4.4 showcases a downstream application on a type ad-hoc WSD task in a specific semantic field (i.e. the concepts MASS and WEIGHT in scientific language) that is common in Humanities reseearch but often overlooked in favor of full-coverage WSD. The paper concludes with a discussion and pointers to future work in Section 5.

2 Method

2.1 Architecture

The present approach deploys a parameter-free architecture which is heavily inspired by both the Matching Networks (Vinyals et al., 2016) and the Prototypical Networks (Snell et al., 2017) frameworks.

In this non-parametric approach, we fine-tune a given LLM using episodic training. In this type of training, each batch constitutes a training episode which is designed in order to match the experimental conditions expected at inference time. In the case of WSD, each episode consists of a number of randomly sampled sentences—a 'support set'—, exemplifying different word senses of a given lemma, as well as a second set of randomly sampled sentences—a 'query set'—, for which a word sense prediction needs to be made.[2]

Sentences in the query set are used in order to obtain a contextualized word embedding for the target word (i.e. the word representing the lemma to be disambiguated). The support set is used in order to compute abstract word sense representations for each of the word senses a lemma may have.[3]

[1] Both models can be accessed through the original `huggingface` repository through the following links https://huggingface.co/emanjavacas/MacBERTh and https://huggingface.co/emanjavacas/GysBERT.

[2] See Table 3 in the Appendix for an illustration of the structure of the lexical databases and an example of the sentences that are being classified.

[3] Note that for this approach to work, the true word sense

These abstract sense embeddings can be computed in multiple ways, but for simplicity we have chosen a centroid approach. First, the contextualized embeddings of the target lemma in the support sentences are extracted, and, then, averaged in order to get a single representation per word sense.[4]

More formally, let $E(S_i)$ and $E(Q_j)$ denote the contextualized embeddings of the target lemma in the i^{th} support sentence and the j^{th} query sentence in the current training episode. And let $R(S_j)$ denote the word sense of the j^{th} sentence in the support set. Then, the representation for the k^{th} word sense r_k in a given training episode is computed as follows:

$$E(r_k) = \sum_{\{j|R(S_j)=k\}} \frac{E(S_j)}{|\{j|R(S_j) = k\}|} \quad (1)$$

The objective of the approach is to maximize the probability of the true word sense given by the following equation.

$$p(k|E(Q_j)) \propto \text{sim}(E(Q_j), E(r_k)) \quad (2)$$

where **sim** is a similarity function in the embedding space. The probability that a given query example belongs to a given word sense is proportional to the similarity between the embedding of the query sentence and the word sense representation. In order to obtain a valid probability distribution, the similarity scores are normalized using the soft-max function.[5]

We fine-tune the entire set of LLM parameters over a number of training episodes. For each episode, we sample lemmas from the training set uniformly—i.e. disregarding lemma frequency—, which has been found to be helpful in order to improve the classification efficacy for low frequency words (Chen et al., 2021). Moreover, we sample a maximum of 10 sentences for the support set and 20 sentences for the query set. Each model is trained for a maximum of 3,000 training steps, or less if convergence is reached, as indicated by development performance.

of all sentences in the query set needs to be represented in the support set.

[4]During all the present experiments, we take the output of the last hidden layer as the contextualized embedding. Moreover, if the target word was sub-tokenized into multiple sub-words, we average over the embeddings of these sub-words.

[5]From this point of view, the present approach resembles metric-based methods in the context of few-shot classification. See also Chen et al. (2021) for an application to WSD.

2.2 Historical Pre-Training

As we target WSD in historical text, the NLP models we employ need to address a number of additional difficulties that are usually not present when dealing with present-day text. First, historical languages often have non-consolidated spelling, which leads to an increased amount of orthographic variation. This is further aggravated by the fact that historical text exists primarily in printed or handwritten form, and hence requires error-prone digitization techniques to be computationally processed. Finally, in many studies, the collection of historical text under scrutiny covers a large time span and, thus, the language used in these texts has been subject to grammatical and semantic change.

Following previous research (Hosseini et al., 2021; Manjavacas and Fonteyn, 2021, 2022), we resort to historically pre-trained LLMs in order as the basis for our WSD experiments. More specifically, we deploy LLMs that are pre-trained from scratch on historical data instead of adapted from present-day models, since the former strategy has been shown to yield stronger performance when applied to historical data (Manjavacas and Fonteyn, 2021, 2022). For English, we used `MacBERTh` (Manjavacas and Fonteyn, 2022), and for Dutch we use the newly introduced `GysBERT`, which will be described in more detail in Section 3.1.

3 Datasets

The datasets underlying the present study come from the Oxford English Dictionary (henceforth: OED Oxford University Press) and the "Woordenboek der Nederlandsche Taal" (Dictionary of the Dutch Language, henceforth: WNT Instituut voor de Nederlandse Taal). Both resources consist of large historical lexicons, where each lemma is categorized into a hierarchy of word senses. Each word sense is given a definition, and is exemplified by a set of sentences spanning a certain time window.

To construct a suitable corpus for testing our WSD approaches, we sampled 1,000 words according to frequency from each language [6] and searched for them in the corresponding resource. The collected data for each language is described in Table 1.

On this data set, we produced a 10% split of lemmas, which were used to evaluate models on

[6]More specifically, we made sure to sample equally from different frequency bands in order to obtain a representative sample of the vocabulary.

	Lemmas	Senses	Quotations
OED	846	22,004	121,684
WNT	755	22,547	137,131

Table 1: Summary statistics of the used datasets.

unseen lemmas. We refer to this as the "held-out set". Finally, for each lemma, we produced a 50% split of quotations, following the original distribution of word senses.

3.1 GysBERT: A Historically Pre-Trained LLM for Dutch

In order to process historical Dutch, we have developed `GysBERT`, a historically pre-trained model for Dutch. To our knowledge, `GysBERT` represents the first such model for Dutch. Architecturally, `GysBERT` closely follows BERT-base uncased. For pre-training purposes, we compiled a data set using two databases of historical Dutch texts.

The first data set is Delpher, a database of historical newspapers, books and journals that comprises more than 130 million scanned and digitized pages, spanning from 1618 to the end of the 20^{th} century (Koninklijke Bibliotheek). The second set is the Digital Library of Dutch Literature (DBNL) (Koninklijke Bibliotheek et al.). The DBNL consists in a comprehensive digital library of Dutch literature that resulted from the joint effort of Dutch and Flemish libraries, and aims to represent the entire linguistic area.

While the Delpher database contains OCR'd text of varying quality, the DBNL is the result of a thorough digitization campaign and presents generally high quality transcriptions. In order to make sure that only text of sufficiently quality is used for pre-training, we developed the following filtering strategy. First, we trained statistical character-level 5-gram language models using `KenLM` (Heafield, 2011). Specifically, we trained a single model per century of text available from the clean DBNL data. Then, for each snippet of Delpher data, we obtain a quality estimate as the perplexity that the corresponding DBNL-based model assigns to it. Manual observation of random snippets suggested discarding texts with a perplexity of 20 or higher. Furthermore, we restricted ourselves to texts published between 1500 and 1950.

In total, the remaining data set consists of 5.8B tokens from Delpher and 1.3B tokens from DBNL—which amounts to ca. 7.1B tokens. We used this data set in order to train a WordPiece tokenizer with a vocabulary of 30,000 tokens, and pre-trained BERT with default parameters, for 1,000,000 training steps, keeping the maximum sequence length at 128 subtokens.[7]

4 Experiments

In order to test the efficiency of this non-parametric approach, as well as the impact of historical pre-training, we ran a series of experiments comparing historically pre-trained models and present-day models. For English, we compare `MacBERTh` to `BERT`—which corresponds to the official release of BERT-base uncased (Devlin et al., 2019). For Dutch, we compare `GysBERT` with the BERT-based `BERTje` (de Vries et al., 2019) and the RoBERTa-based (Liu et al., 2019) `RobBERT` (Delobelle et al., 2020).

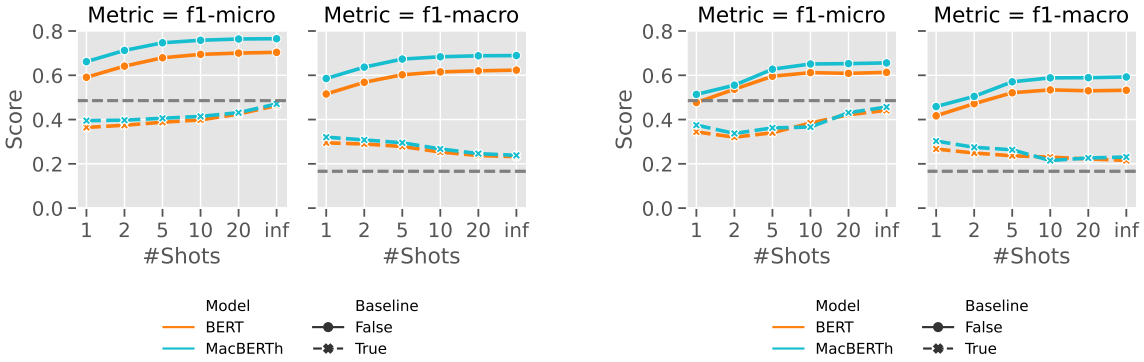
Moreover, for each model, we compare results with respect to non fine-tuned versions of the models—we refer to these variants as frozen baselines. When applicable, we also report the results obtained by a most frequent sense (MFS) baseline. We focus on F1-scores averaged over the different lemmas. Since the distribution of word senses is often heavily skewed, we report both micro and macro F1-scores.

4.1 General Results

Figure 1 and Figure 2 show the average F1-scores over **in-domain lemmas**—i.e. these lemmas were present in the training data, even though the specific sentences on which these results are computed were absent—and **held-out lemmas** for the different models in the full training data regime. In these plots, we highlight the effect of using an increasing number of shots (shown on the x-axis). Note that the number of shots in this context refers to the number of available support examples for each word sense during inference.

Figure 1 and Figure 2 show that the proposed fine-tuning approach is very efficient with respect to the frozen baselines, as we observe an increase of 0.2 points or more. The effect is larger when considering macro F1-scores and an increase in the number of shots, indicating that the proposed fine-tuning yields more discriminative models ir-

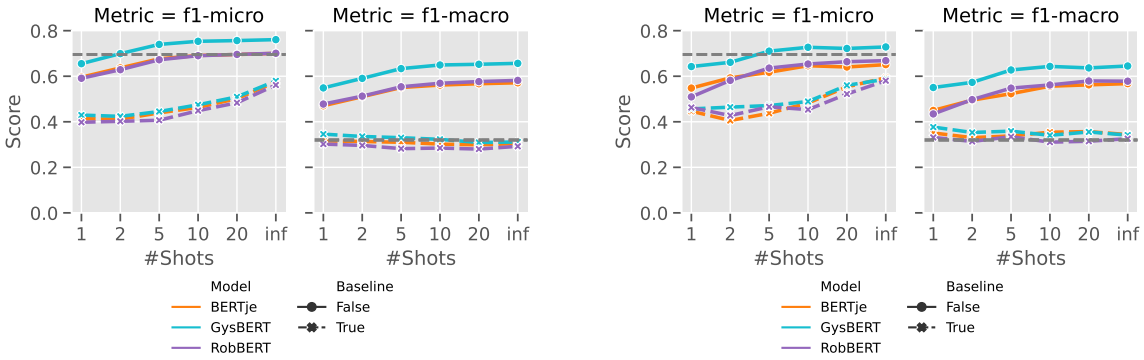
[7] `GysBERT` will be released on the `huggingface` platform.



(a) In-domain

(b) Held-out

Figure 1: Results of WSD on the OED for in-domain lemmas (a) and held-out lemmas (b). Solid lines denote the proposed models trained on the full data sets. Dashed lines represent the corresponding frozen baselines. The MFS baseline is shown by a grey dashed line. The x-axis (number of shots) corresponds to the number of example sentences per sense shown during inference.



(a) In-domain

(b) Held-out

Figure 2: Results of WSD on the WNT for in-domain lemmas (a) and held-out lemmas (b). Solid lines denote the proposed models trained on the full data sets. Dashed lines represent the corresponding frozen baselines. The MFS baseline is shown by a grey dashed line. The x-axis (number of shots) corresponds to the number of example sentences per sense shown during inference.

respectively of the skewness of a given lemma's word senses. Interestingly, this happens under both in-domain and held-out conditions. Overall, frozen baselines underperform the MFS baseline in terms of micro F1. In terms of macro F1, frozen baselines do outperform the MFS baseline, albeit very slightly.

For held-out lemmas, there is just a mild decrease in performance, of less than 0.1 points in both data sets. This indicates encouraging generalization capabilities. Focusing on the x-axis, we observe that an increase in the number of shots results in a continuous increase in performance up until 5 shots, after which the improvement plateaus.

Overall, performance is slightly higher in the Dutch data set, which may be due to a larger skewness in the distribution of word senses. This skewness can, indeed, be inferred from the high micro F1-score obtained by the MFS baseline.

Finally, historically pre-trained models outperform present-day models with a safe margin. This result is particularly relevant in our case, since the superiority of historical pre-training cannot be concluded on the basis of the frozen baselines alone, but surfaces only after applying the non-parametric fine-tuning.

4.2 Small Training Regime

In order to assess the efficiency of the proposed approach on the low-data regime, we performed a series of experiments in which both the number of lemmas and the maximum number of examples per

sense are limited **during training**. The results are shown in Figure 3 for the OED and in Figure 4 for the WNT. For inference, we keep the number of shots at 5. Note that the performance in these experiments refers to inference on held-out lemmas.

In Figure 3, we observe that the historically pre-trained models are consistently more effective than the present-day counterparts across training conditions. Furthermore, we observe that even a very small amount of training data (e.g. 50 training lemmas in total) yields consistent gains over the frozen baselines, regardless of the number of examples per sense.

Moreover, the effect of number of lemmas is small when using only 2 or 5 examples per sense. When using 10 or 50, an increase in the number of lemmas has a positive effect on performance up to 500 lemmas. Doubling this amount to 1,000 lemmas, however, yields little return. These experiments seem to indicate that strong generalization can be achieved with relatively small training data sets (e.g. 500 training lemmas and 10 example sentences per sense).

In the case of the Dutch data set from Figure 4, we observe similar patterns to those from the OED data set. Again, micro F1-scores are very high for the MFS baseline, and a larger number of training lemmas (i.e. 500) and number of examples per sense (i.e. 50) are needed in this data set for the models to outperform the MFS baseline.

4.3 Impact of Time

Since the example sentences of both the OED and WNT data sets display the publication year of the work in which they appear, we can inspect the performance of the different models over sentences in different time periods. From this angle, we expect to observe an improvement in performance for the earlier periods when the fine-tuned model was pre-trained historically. Figure 5 shows the time-aggregated results with the century on the x-axis.

The historically pre-trained models outperform the present-day models across the entire range. Moreover, these plots confirm that the relative improvement over present-day models is indeed larger in the earlier centuries, where the challenges presented by historical text are most acute.

4.4 Downstream Application

So far, we have examined the performance of the non-parametric fine-tuning on the basis of the lexical databases (OED and WNT), which offer large quantities of available training data and allow us to control the training conditions. In order to test the efficiency of non-parametric fine-tuning on smaller-scale scenarios, which can be considered more 'realistic' in the context of Humanities research, we ran an experiment on a classification task involving an ad-hoc WSD task around the word senses of the lemmas *mass* and *weight* in 18th and 19th century scientific writing.

This experiment is part of on-going research aimed at tracing the development of the concept of MASS when Newtonian physics forced a process of semantic differentiation between the terms *mass* and *weight*. To this end, all 56,813 instances of *mass* and *weight* in the Royal Society Corpus (RSC Fischer et al., 2020) will be analyzed with respect to a fine-grained classification of 6 word senses—see A for examples of these categories. With the goal of automating the annotation process, a sample of 1,500 instances—including 621 cases of *mass* and 879 of *weight*—was first manually annotated by a domain expert.

Subsequently, we set up a total of 4 competing fine-tuning approaches, including the non-parametric approach described in Section 2, and 3 additional ones to serve as baselines. The first one, `Standard`, consists in fine-tuning a classification layer on top of `MacBERTh`, as implemented in the `transformers` library (Wolf et al., 2020). The remaining two involve a K-Nearest Neighbours (`KNN`) and a Support Vector Machine classifier (`SVC`) on top of the token-embeddings produced by `MacBERTh`.[8] We optimize the models using a 10-fold Cross Validation procedure, where each fold respects the original word sense proportions.[9] For the non-parametric fine-tuning approach, we follow the hyper-parameterization from the main experiments reported in this paper.

We focus on micro and macro F1-scores, reporting means and standard deviations for each model. The results are shown in Table 2. The non-parametric approach outperforms the baselines in terms of both micro and macro F1-scores. However, taking into account the standard deviation from the Cross Validation, the advantage with respect to the best baseline in terms of macro F1-score does not hold.

[8]The latter two baselines were implemented with the `scikit-learn` library (Pedregosa et al., 2011).

[9]In the case of the `KNN` classifier, we hyper-optimize the number of nearest neighbors, as well as the distance metric. In the case of `SVC`, we hyper-optimize the C parameter.

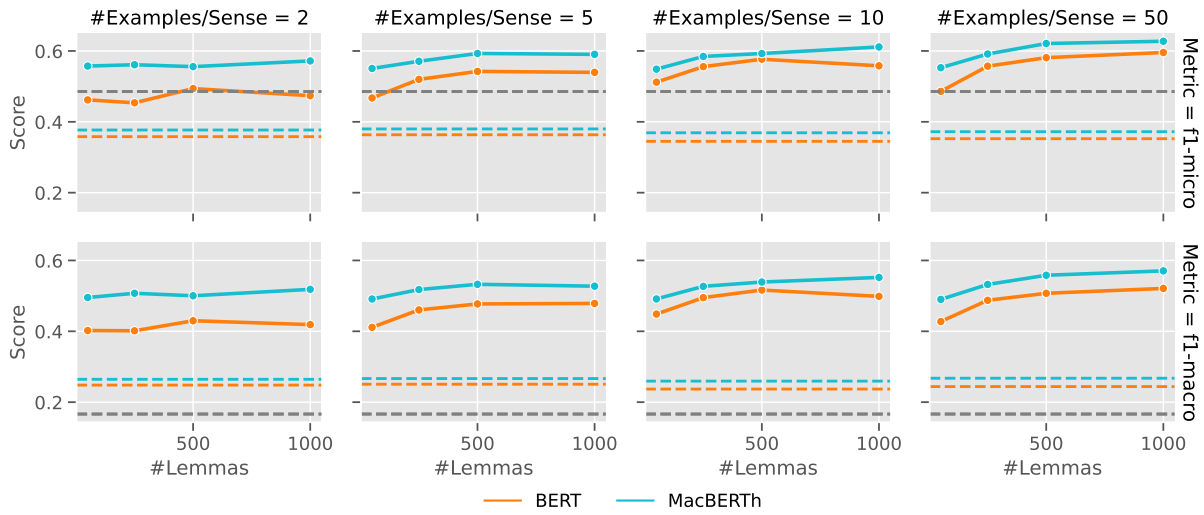


Figure 3: F1-scores for WSD on OED held-out lemmas for the proposed models trained on 50, 250, 500 and all lemmas (on the x-axis) and 2, 5, 10, and 50 example sentences per sense (on the columns).

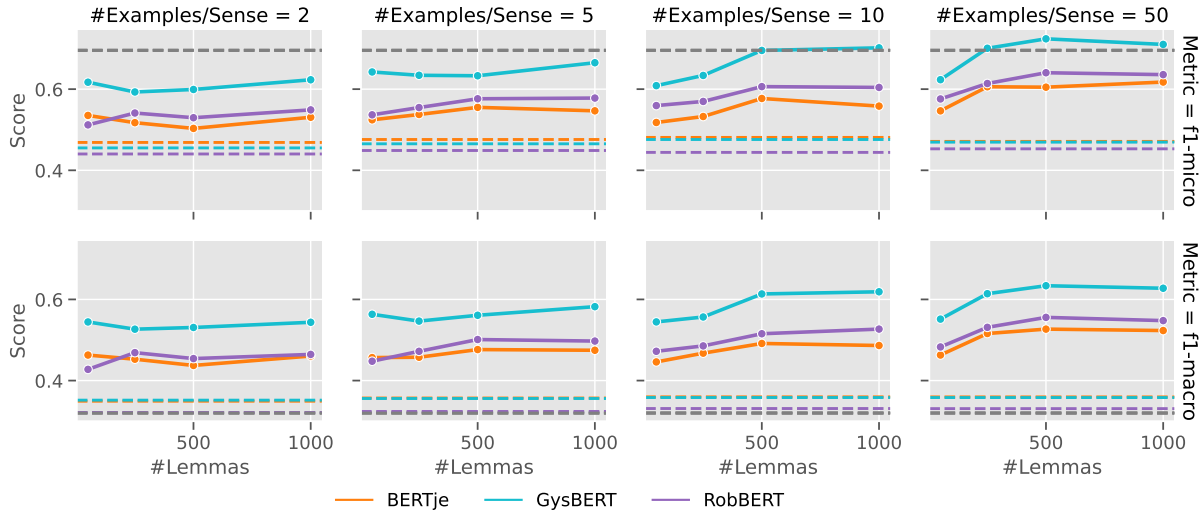


Figure 4: F1-scores for WSD on WNT held-out lemmas for the proposed models trained on 50, 250, 500 and all lemmas (on the x-axis) and 2, 5, 10, and 50 example sentences per sense (on the columns).

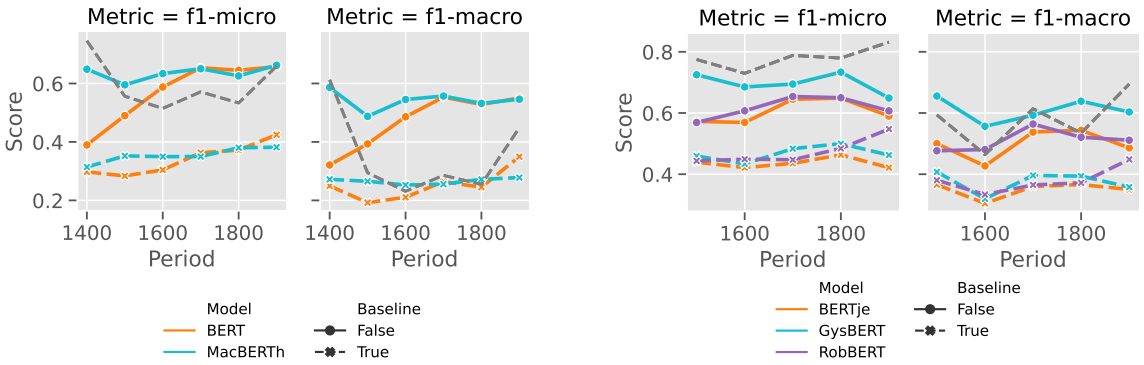
Surprisingly, the standard fine-tuning approach is not only less competitive than the baselines, but also suffers from strong variance across CV folds. This is probably due to the small number of training examples available for fine-tuning, and the large number of parameters that need to be tuned in this approach. In contrast, the non-parametric approach achieves not only the highest scores but also the lowest standard deviation of all competitors, indicating that this may be a much better suited approach for fine-tuning on small training data sets.

	Micro F1		Macro F1	
Model	Mean	StdDev	Mean	StdDev
KNN	0.830	0.007	0.695	0.032
SVC	0.819	0.007	0.601	0.035
Standard	0.827	0.029	0.520	0.066
Non-Parametric	**0.864**	0.006	**0.699**	0.025

Table 2: 10-fold cross-validated results of the classification experiments of "mass" and "weight" for 4 different fine-tuning methods. Best performing result in **bold**.

5 Discussion & Future Work

Our experiments highlight that Humanities researchers who seek to automatically disambiguate word senses over time may be able to do so with reasonable performance, even when they provide only a small amount of sentences exemplifying the target word senses and/or leverage general-purpose



(a) OED

(b) WNT

Figure 5: Results of WSD over different periods of time (on the x-axis), using the fine-tuned models (solid lines) as well as the corresponding frozen baselines (dashed lines) and an MFS baseline (grey dashed line). Inference was done on held-out lemmas using 5 shots and the full-training regime.

lexical resources, such as the OED or the WNT. More specifically, in the full training data scenario, held-out lemmas could be classified with micro F1-scores of 0.627 for English and 0.71 for Dutch, using the historically pre-trained models. These results imply 40.3% and 34.22% improvements over the respective frozen baselines. Moreover, we observe that even a small number of training lemmas can lead to important improvements over frozen baselines. For example, when training on just 50 lemmas and only 2 instances per sense, micro F1-scores can be obtained of 0.557 for English and 0.617 for Dutch, which represent improvements of 32.8% and 24.4% over the frozen baselines.

In this sense, we go a step further than Chen et al. (2021), who—in contrast to our experiments—leveraged the training data in order to construct word sense representations at inference time. By doing so, they assumed that all lemmas in the test data are known from training data. What we found is that the fine-tuning approach is also effective on held-out lemmas, which means it can be applied in cases where practical constraints exist on the amount of available annotated data.

Our experiments also highlighted that historically pre-trained models are able to better handle the intricacies of historical data sets than present-day models when applying the discussed non-parametric fine-tuning approach. This result is particularly important in the present context, since the superiority of historical pre-training is not apparent on the basis of the frozen embeddings only. Using the frozen embeddings, the difference in performance between the historically pre-trained models and the present-day models is negligible. Moreover, we presented a case study which highlighted that non-parametric fine-tuning can be much more efficient than the more commonly used standard fine-tuning approaches, especially in small training regimes.

The main objective of the present fine-tuning approach is to push the embeddings of the query sentences closer to the non-parametric representations of the true word senses. By conjecture, the proposed approach works by learning to distill the semantic features in the input sentences that are most relevant to lexical semantics, stripping off irrelevant information for WSD. Thus, the fine-tuned model is allegedly able to achieve improved performance when classifying lemmas that have not been encountered during training.

Finally, we wish to note that we limited ourselves to normalized dot products as the measure of relatedness between representations in this study, and we deployed the transformer (Vaswani et al., 2017) architecture underlying BERT as is. Future work could, however, investigate what can be gained by experimenting with other similarity functions, and adding more complex layers such as an attention module over different word sense representations.

References

Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2751–2761, Online. Association for Computational Linguistics.

Andreas Blank. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In Andreas Blank and Peter Koch, editors, *Historical Semantics and Cognition*, page 61–90. Mouton de Gruyter, Berlin/New York.

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

Howard Chen, Mengzhou Xia, and Danqi Chen. 2021. Non-parametric few-shot learning for word sense disambiguation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1774–1781, Online. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based language model. *arXiv preprint arXiv:2001.06286*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yingjun Du, Nithin Holla, Xiantong Zhen, Cees Snoek, and Ekaterina Shutova. 2021. Meta-learning with variational semantic memory for word sense disambiguation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5254–5268, Online. Association for Computational Linguistics.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The royal society corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 794–802, Marseille, France. European Language Resources Association.

Lauren Fonteyn. 2020. What about grammar? Using BERT embeddings to explore functional-semantic shifts of semi-lexical and grammatical constructions. *CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands*, pages 257–268.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. *arXiv:2004.14118 [cs]*. ArXiv: 2004.14118.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, China. Association for Computational Linguistics.

Thora Hagen, Erik Ketzan, Fotis Jannidis, and Andreas Witt. 2020. Twenty-two historical encyclopedias encoded in TEI: a new resource for the digital humanities. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 112–120, Online. International Committee on Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Learning to learn to disambiguate: Meta-learning for few-shot word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4517–4533, Online. Association for Computational Linguistics.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. Neural Language Models for Nineteenth-Century English (dataset; language model zoo).

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Instituut voor de Nederlandse Taal. Woordenboek der nederlandsche taal. `https://gtb.ivdnt.org/search/`.

Koninklijke Bibliotheek. Delpher. https://www.delpher.nl/.

Koninklijke Bibliotheek, Taalunie, and howpublished="https://www.dbnl.org/" Vlaamse Erfgoedbibliotheken, title=Digitale Bibliotheek voor de Nederlandse Letteren (DBNL).

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. *ArXiv*, abs/1907.11692.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018. Incorporating glosses into neural word sense disambiguation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2473–2482, Melbourne, Australia. Association for Computational Linguistics.

Enrique Manjavacas and Lauren Fonteyn. 2021. Macberth: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on NLP4DH @ ICON 2021*, pages 23–36, online.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs pre-training language models for historical languages. *Journal of Data Mining and Digital Humanities*.

Oxford University Press. Oxford english dictionary. https://www.oed.com/.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis lectures on human language technologies*, 5(2):1–157.

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.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with Latent Semantic Analysis. In Kathryn Allan and Justyna A. Robinson, editors, *Current Methods in Historical Semantics*. De Gruyter, Berlin, Boston.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Nina Tahmasebi, Lars Borin, Adam Jatowt, et al. 2018. Survey of computational approaches to diachronic conceptual change. *arXiv preprint arXiv:1811.06278*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

lemma	sense	definition	example	year
RAT	1.a.	Any rodent of the genus Rattus and related genera of the family Muridae, resembling a large mouse, often with a naked or sparsely haired tail.	"Rats and mice purloin our grain" *J. Gay, Fables, II. viii. 74*	1732
	4.a	A dishonest, contemptible, or worthless person.	"No Female Rat shall me deceive, nor catch me by a crafty wild." *in Roxburghe Ballads VI. 106*	1656
RAT	1.	Knaagdier behoorende tot het geslacht Rattus van de familie der Muridae of 'ware muizen'.	"Daar 't katje woond, daar word het huis Gezuiverd van de Rat, en Muis" *Luyken, Besch. d. W. 223*	1708
	2.	Oneig. toegepast op personen. Armoedzaaier, gelukzoeker.	"Dien grootmaecker, die cael Rat" *Ogier, Seven Hoofts. 19*	1644

Table 3: Example lemma and sense with definition and quotation from the Oxford English Dictionary (top row) and the Woordenboek der Nederlandsche Taal (bottom row).

label	sense	example from Royal Society Corpus
N	*mass* or *weight* refers to thing or object	"The mass on the filter was treated with boiling alcohol" (Edward Schunk, 1853) "a flat circular weight nicely turned, and pierced in the direction of its diameter to receive the bar, was slid upon it" (Henry Kater, 1819)
M	*mass* or *weight* refers to MASS (i.e. how much matter is within an object)	"We are thus led to inquire how the stresses are distributed in the earth 's mass and what are their magnitudes" (G. H. Darwin, 1882) "In the third, the weight of the principle bones of a selected number of species (27) is stated" (John Davy, 1865)
W	*weight* refers to WEIGHT (i.e. referring to force, balancing, counterpoises, or the amount of effort required to lift something)	"fig. 3 is only 40 feet from the bow, and that the excess of weight over buoyancy on this length is only 45 tons" (E.J Reed & G.G Stokes, 1871)
W/M	unclear whether the example refers to MASS or WEIGHT	"The Commissioners for the Restoration of the Standards of weight and measure, in their Report dated December 21, 1841, recommended that..." (W.H Miller, 1856)
COL	*mass* or *weight* refers to a collection of objects (e.g. a mass of small fragments)	"A glacier is not a mass of fragments" (James Forbes, 1846)
MET	*mass* or *weight* is used to indicate the importance of a thing (e.g. the weight of authority)	"The next thought is that I may have assigned too great a mass to the doubt" (John Henry Pratt, 1855) "The contact theory has long had possession of men 's minds, is sustained by a greatweight of authority" (Michael Faraday, 1840)

Table 4: Classification scheme of *mass* and *weight* instances retrieved from the Royal Society corpus. In total, 1,500 examples were manually classified in one of these 6 custom categories.