Adapting vs. Pre-Training Language Models for Historical Languages

Enrique Manjavacas[1] **and Lauren Fonteyn**[1]

[1]Leiden University, The Netherlands

Corresponding author: Enrique Manjavacas , enrique.manjavacas@gmail.com

Abstract

As large language models such as BERT are becoming increasingly popular in Digital Humanities (DH), the question has arisen as to how such models can be made suitable for application to specific textual domains, including that of 'historical text'. Large language models like BERT can be pre-trained from scratch on a specific textual domain and achieve strong performance on a series of downstream tasks. However, this is a costly endeavour, both in terms of the computational resources as well as the substantial amounts of training data it requires. An appealing alternative, then, is to employ existing 'general purpose' models (pre-trained on present-day language) and subsequently adapt them to a specific domain by further pre-training. Focusing on the domain of historical text in English, this paper demonstrates that pre-training on domain-specific (i.e. historical) data from scratch yields a generally stronger background model than adapting a present-day language model. We show this on the basis of a variety of downstream tasks, ranging from established tasks such as Part-of-Speech Tagging, Named Entity Recognition and Word Sense Disambiguation, to ad-hoc tasks like Sentence Periodization, which are specifically designed to test historically relevant text processing.

Keywords

Natural Language Processing;Historical NLP;Language Models

I INTRODUCTION

In recent years, there has been a great interest within the Digital Humanities research community to utilize semantic vector representation algorithms for the computer-aided retrieval, annotation, and analysis of textual data. By means of such algorithms, the contextual distribution of linguistic items – which could be words, but also phrases, sentences, or even longer chunks of text – can be represented as (compressed) numeric vectors that serve as a proxy of their meaning [e.g. Turney and Pantel, 2010, Erk, 2012, Lenci, 2018]. As these semantic vector algorithms grew in popularity, it soon became clear that they could be of service to the Digital Humanist in various ways. By training a computational model to generate a vector representation – or 'embedding' – for a given word, the Digital Humanist can for instance use these embeddings to automatically retrieve documents that discuss the concept the word refers to without naming it explicitly [Wevers and Koolen, 2020], or to extract all semantically related words (e.g. *friendship*: near-synonyms, *amity*; antonyms, *hostility*) from a target text collection [e.g. van Eijnatten and Ros, 2019, Ehrmanntraut et al., 2021]. Moreover, it has been shown that such word embeddings may also be used as a data-driven means of revealing gender bias in textual material [e.g. Wevers, 2019], mapping character relations in novels [e.g. Grayson et al., 2016], and, when applied to diachronic text collections, detecting changes in word meaning over time [Sagi et al., 2011, Tahmasebi et al., 2018, Kutuzov et al., 2018, Sommerauer and Fokkens, 2019, Marjanen et al., 2019, Martinez-Ortiz et al., 2019]. Furthermore, for research questions where word polysemy (when a word has multiple related senses, e.g. *foot* 'body part' and 'base/lowermost part') and homonymy (when a character string has multiple, unrelated meanings, e.g. *bark* 'outer layer of tree trunk', 'sound made by dog') could create methodological issues, researchers have also been catered to by models equipped to create contextualized vector representations for individual word tokens. Such token-based models have proven helpful in word-level tasks, such as sense disambiguation [Fonteyn, 2020, Beelen et al., 2021], Named-Entity Recognition Labusch et al. [2019], Konle and Jannidis [2020], Schweter and Baiter [2019], Schweter and März [2020], Ehrmann et al. [2020a], Boros et al. [2020], Brandsen et al. [2021], Ehrmann et al. [2021], and the automated detection of semantic narrowing (i.e. when a word loses one or more senses and becomes more restricted in it usage) or broadening (i.e. when a word gains one or more new senses and becomes more varied in its usage) [Sagi et al., 2011, Giulianelli et al., 2020]. Moreover, these models can easily be adapted to perform a plethora of higher-level downstream tasks with great accuracy, including automated text classification [Adhikari et al., 2019, Jiang et al., 2021], text segmentation [Pagel et al., 2021] or event detection [Sims et al., 2019], to name a few.

In part, the appeal of semantic vector representation algorithms lies in their potential to automate semantic (rather than formal) data retrieval and annotation, which helps increase the amount of data that can be processed by researchers. At the same time, the application of large language models in humanities research – and in particular, humanities research that focuses on the interpretation and analysis of historical text – may also offer a more objective means to analyse textual data [e.g. Sagi et al., 2011]. Researchers who interpret and analyse historical textual material are well-aware that the interpretation of historical textual material must not be approached with present-day intuitions [Tahmasebi and Risse, 2017]: because there are substantial differences between the way in which concepts and discourses of class, gender, norms and prestige are linguistically represented in different time periods, present-day intuitive judgments of historical language are likely to lead to inaccurate, 'anachronistic' interpretations of the data. A methodological set-up where a computational language model 'substitutes' the manual involvement of the present-day analyst, then, is an attractive approach, as it helps minimize (or even eliminate) such potentially biased, intuitive judgments in the process of data annotation and analysis.

Of course, processing historical text poses a series of challenges for vector representation algorithms, but these may be overcome by large token-based language models. Historical text involves, for instance, high degrees of orthographic variation. This is especially true in the case of Western European languages, which only acquired their modern spelling standards roughly around the 18^{th} or 19^{th} centuries. This introduces a 'layer of variation' that type-based model will struggle with: an algorithm that produces word vectors for each unique string of letters will not conflate the contextual distribution of *remembring* and *remembering*, despite the fact that these are spelling variants of the same word type. Secondly, historical text requires digitization before it can be processed by computational means, but current OCR and HTR technology is error-prone, and manual correction is costly. As a result, digitized historical text often involves an additional layer of variation, which, in contrast to orthographic variation, is characterized by near-random distributions (and hence difficult to reduce). Yet, due to the enhanced capacity to leverage context of large, token-based language models, they should be able to abstract over the mentioned layers of variation and produce more accurate semantic representations than their type-based counterparts.

Furthermore, the incipient paradigm-shift associated with the dawn of large language models has also proven appealing because it can bring infrastructural advantages to less tech-savvy communities. The new paradigm – known as the 'pre-train-then-fine-tune' paradigm – involves first preparing a language model by means of 'pre-training' on a large background collection of text, and then 'fine-tuning' this language model in order to perform a particular task. Pre-training typically involves very large collections but no annotation, and once pre-trained, fine-tuning necessitates a manually labeled dataset exemplifying the target task. Importantly, thanks to their high "sample efficiency" [Kaplan et al., 2020], large language models can achieve significant performance on the basis of comparatively smaller training datasets than those required by other Machine Learning architectures. This allows for an advantageous collaboration model, in which tool developers focus on producing high quality language models, and humanities researchers focus on creating annotations for the desired task, on which these language models can be fine-tuned following standard procedures.

Still, despite these positive notes, researchers who want to call upon these language models to target historical text will face practical hurdles. In particular, the training of large models such as BERT is a costly endeavour, both in terms of computational resources, as well as the substantial amounts of training data it requires. For historical data, digitized corpora are often exhaustive, but still small. As a result, past work with historical data has resorted to employing pre-fab, present-day language models [Giulianelli et al., 2020, Hosseini et al., 2021a], which are occasionally adapted for historical usage by further pre-training on historical datasets. Unfortunately, such a set-up may be problematic for several reasons.

First, current large language model implementations rely on tokenization procedures – like Byte-Pair-Encoding (BPE) [Gage, 1994, Sennrich et al., 2016] or WordPiece [Schuster and Nakajima, 2012] – that break down character sequences into so-called sub-word tokens, optimizing a particular information-theoretic measure. Originally, the motivation for this tokenization approach was tackling the out-of-vocabulary word problem. For any given character sequence that was not included in the training set, traditional approaches would struggle to generate a vector representation, given that no vector was assigned to it in the original vector space. With the current tokenization approach, a vector representation is computed through composition of the vector representation of the sub-word tokens into which the original string is decomposed. However, since the adaptation of a pre-existing model implies a tokenizer that has been optimized on present-day language data, the application of such a model on historical data may result in uninformative sub-word tokenizations and ultimately in out-of-domain sequences of sub-word tokens for which the model can only generate low quality vector representations.

Second, it could be problematic to employ a model trained on data that may import the problematic, anachronistic biases towards grammar and semantics that researchers are trying to avoid. In fact, some classification error analyses of historical sense disambiguation tasks indicate that present-day language models indeed erroneously impose a present-day interpretation onto the historical material [e.g. Fonteyn, 2020]. Intuitively, one would expect that pre-training on present-day data (for which there is no issue of sparsity) would provide an advantageous starting point, assuming that large parts of the core grammar and lexical semantics of a language have stayed constant over time. However, the reservations on imported biases and the aforementioned fixed tokenization schemes may run counter to any observations that present the adaptation of pre-existing models as the more data-efficient alternative.

In this paper, we investigate which of the two strategies is bound to produce higher quality vector representations: (i) pre-training from scratch or (ii) adapting a pre-existing model. To this end, we extend the experiments of previous work on pre-training MacBERTh – a large historical language model for English [Manjavacas and Fonteyn, 2021]. More specifically, we compare the performance of a number of models representative of both methods (pre-training vs. adaptation) on a number of downstream tasks. Our experiments are consistent with previous results, which highlight that pre-training from scratch may be a better strategy. The results also suggest that the fixed tokenization models that are currently in vogue may be a bottleneck in the process of successfully adapting large language models.

II MODEL OVERVIEW

While there is a variety of model architectures that could be used for historical NLP, the present study relies on BERT, a stack of transformer layers with a self-attention mechanism [Vaswani et al., 2017] that optimize a Masked Language Model (MLM) objective [Devlin et al., 2019]. Despite the existence of several MLM alternatives, our choice to work with BERT is motivated by the fact that (i) it is well-established and thoroughly studied, and that (ii) the on-going evaluation of alternative choices—mostly focused on Natural Language Understanding (NLU) tasks—has not yielded a clearly superior architecture and (iii) experimenting with alternative architectures involves a multiplicative cost factor on the pre-training and fine-tuning part of the experiments presented in the current study, which would, unfortunately, surpass the available budget.

In order to quantify the relative advantage of the alternative pre-training methods for historical language, we compare the following instantiations of BERT.

The first BERT model we consider is trained on **present-day English** data only, corresponding to "BERT-Base Uncased" in the original repository. This model, to which we will henceforth refer as `BERT`, is trained on ca. 3.3B tokens—i.e. the BookCorpus [Zhu et al., 2015] and the English Wikipedia—using a WordPiece [Schuster and Nakajima, 2012] vocabulary of 30,000.

Second, we consider two variants of `BERT`—i.e. "BERT-Base Uncased"—which are subsequently adapted by **further pre-training on historical English data**. The first 'historically adapted' model we consider has been fine-tuned at the Alan Turing Institute on 5.1B tokens of historical English text published between 1760 and 1900 [Hosseini et al., 2021a] .[1] We will refer to this model as `TuringBERT`.

Considering the relatively limited time span covered by `TuringBERT`, we also created a second adapted model, which we will refer to as `BERT-Adapted`. This model also corresponds to an instantiation of "BERT-Base Uncased", but, in this case, it is further pre-trained on the same historical collection that served as the basis for developing MacBERTh [Manjavacas and Fonteyn, 2021]. This collection contains a large sample of English text covering a time span from 1473 to 1950, and includes the Early English Books Online (EEBO) corpus (1473-1700), the Evans Early American Imprints Collection (EVANS; 1639-1800), Eighteenth Century Collections Online (ECCO; 1701-1800), the Corpus of Late Modern English Texts (CLMET3.1; 1710-1920), the Hansard corpus (Hansard; 1803-1950), and the Corpus of Historical American English (COHA; 1810-1950). The resulting corpus has a total size of ca. 3.9B (tokenized) words, covering a varied range of text types, including literary works, religious and legal text, parliamentary debate transcriptions, as well as news reports and magazine articles. The pre-processing procedure involved the removal of foreign text, for which we used an ensemble of the Google's Compact Language Identifier (v3) and the FastText Language Identification system [Grave, 2017], operating over chunks of 500 characters, which were flagged as foreign whenever both systems indicated a language other than English as the highest probability

[1] The model is available through the accompanying online repository [Hosseini et al., 2021b].

language. Subsequently, the text was split into sentences using the NLTK built-in sentence tokenizer [Bird, 2006].

Finally, we consider the performance of the Present-day and historically adapted models in light of the **historically pre-trained** model MacBERTh [Manjavacas and Fonteyn, 2021]. For the creation of MacBERTh, we relied on the seminal implementation of BERT, [2] with the hyper-parameterization corresponding to the "BERT-base Uncased" architecture.[3] Pre-training was done with default parameters, except for the maximum sequence length (set to 128 subtokens) for 1,000,000 training steps. A summary of all compared models can be found in Table 1.[4]

Model	Source	Historical	Adapted	Training Data	Time Span	Vocabulary
BERT	BERT-base Uncased	✗		3.3B		30,000
TuringBERT	BERT-base Uncased	✓	✓	5.1B	1760-1900	30,000
BERT-Adapted	BERT-base Uncased	✓	✓	3.9B	1450-1950	30,000
MacBERTh		✓	✗	3.9B	1450-1950	30,000

Table 1: Overview of all the models involved in the present experiments.

III EXPERIMENTS

In order to assess the relative merit of the alternative approaches, we put the four competing models through a set of downstream evaluation tasks. These tasks were selected on the basis of their relevance for historical text processing.

3.1 Part-Of-Speech Tagging

The first task we consider is Part-Of-Speech (POS) Tagging for Historical English. A particularly suited dataset for our evaluation purposes is the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME) Kroch et al. [2004]. The PPCEME consists of a collection of Early Modern English letters (time span: 1450-1700) that have been annotated manually with morphological and syntactic information. The collection is divided in 448 individual documents, and comprises approximately 1.7M words.

In order to accomplish POS-tagging, a language model is fine-tuned to perform token-level predictions over the input sequence. For each input token, the vector representation for that token is used as features in order to perform classification over the set of possible output POS-tags.[5] For all experiments, we replicate the training and test splits from Han and Eisenstein [2019], which reserve a total of 115 files for testing and from the remaining 333 uses 316 for training and 17 randomly sampled files (ca. 5%) for development.

One of the expected advantages of robust pre-trained language models is their so-called "sample efficiency" – or the ability to generalize from comparatively smaller amounts of training data. In order to compare the candidate models from this perspective, we also run a series of experiments in which we increase the size of the training data, starting at just 50 files up until we reach the full training set of 316 files. For evaluation purposes, we compute accuracy for over all tokens,

[2] Available on the following URL: https://github.com/google-research/bert.

[3] See the original paper [Devlin et al., 2019] for a description of these parameters.

[4] MacBERTh itself is available through the HuggingFace hub: https://huggingface.co/emanjavacas/MacBERTh.

[5] Due to sub-word tokenization, several vector representations may be available for a single input token if this has been split. In those cases, we follow a strategy that ignores all but the first sub-word token.

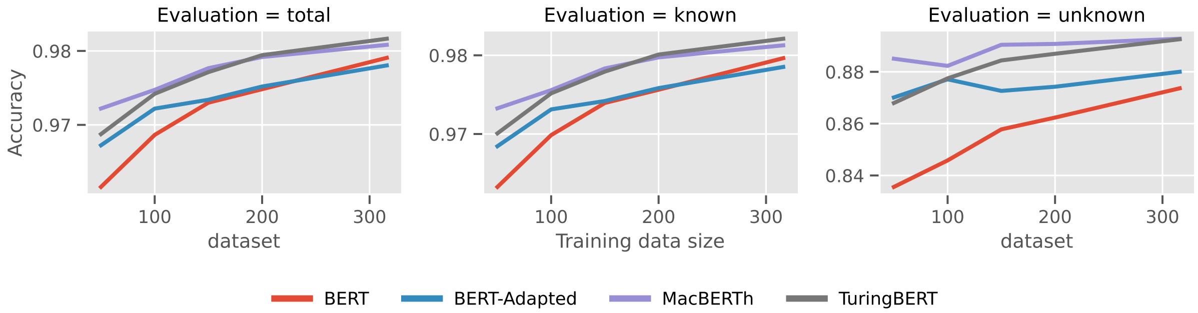

Figure 1: Line plots assessing the sample efficiency of the different candidates. The x-axis represents the number of training files, and the y-axis represents the accuracy. The evaluation is further divided into total, known and unknown tokens, depending on whether input tokens were seen during training or not.

tokens that were seen during training (known), and tokens that were not seen during training (unknown).

Figure 1 shows the results of this experiment. Here, MacBERTh shows peak performance across all conditions, with the difference being larger in the lower data regime. TuringBERT follows up and equals MacBERTh in higher data regimes (starting with 100 files in the training set). Further down the ranking, we find BERT-Adapted, which overall shows lower performance than the other adapted model TuringBERT – except in the lower data regimes when considering unknown tokens. Finally, BERT-Adapted surpasses its non-adapted variant BERT, except for the higher data regimes. It is interesting to note that BERT-Adapted shows considerably worse performance than TuringBERT. As noted in the introduction, these two models were adapted from the same present-day model, and differ only on the underlying historical dataset used for pre-training. However, in light of the performance obtained by MacBERTh, the historical pre-training dataset underlying BERT-Adapted would be expected to provide a stronger model – but this does not seem to be the case.

Finally, we inspect the performance of the different models as a function of the the period from which the target sentences stem. In order to simplify the visualization, we compute accuracy offsets of the different models with respect to the BERT baseline. The results are shown in Figure 2.

Again, the largest performance advantages of MacBERTh are located in the lower training data regimes. Factoring in the period of the target sentence, we also observe that the advantage is not restricted to the time periods to which only MacBERTh and BERT-Adapted had access – i.e. the earlier periods – but also appears in the later periods. Furthermore, it appears that the adapted models, TuringBERT and BERT-Adapted, behave in a largely similar manner. However, their behaviour deviates in the larger training data regimes, where BERT-Adapted's performance dips considerably, eventually underperforming even the present-day BERT baseline on known tokens.

3.2 Named Entity Recognition

The second task we approach is Named Entity Recognition (NER) in Historical texts. We use the dataset provided for the second iteration of the CLEF-HIPE (Named Entity Processing in Historical Newspapers) shared task [Ehrmann et al., 2020b]. For this iteration, the organizers proposed two major tasks – Named Entity Recognition and Classification (NERC) and Entity Linking (EL) [Ehrmann et al., 2022] – covering 5 languages across 6 datasets.

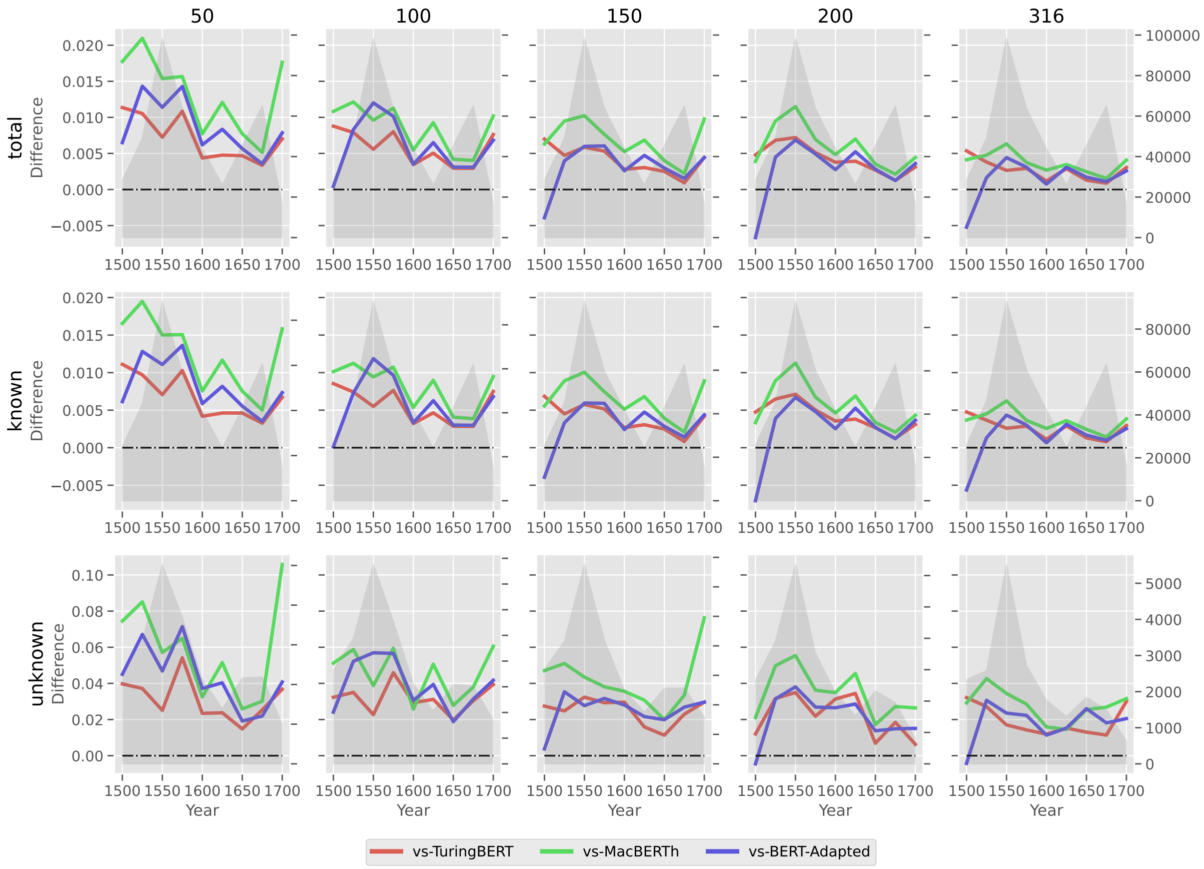

Figure 2: Difference in part-of-speech tagging accuracy in known, unknown and all tokens of all historical models with respect to the present day baseline `BERT`, across different sizes of training datasets.

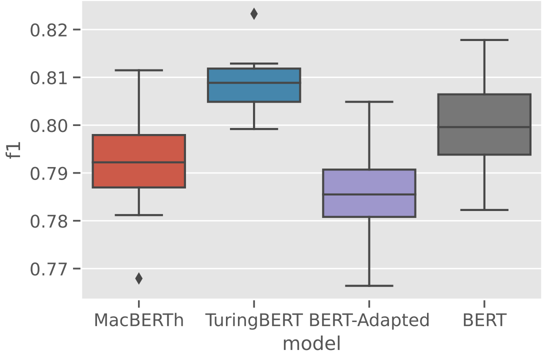

Figure 3: Evaluation in terms of F1-score on the HIPE2022 dataset for Named Entity Recognition in English.

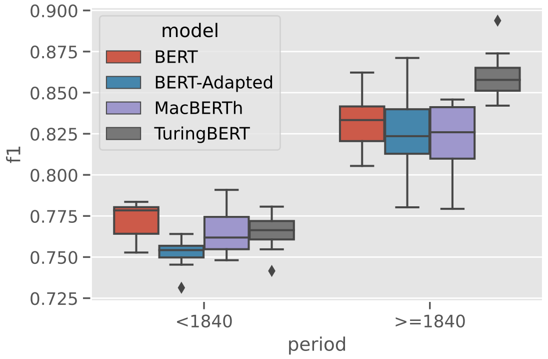

Figure 4: Evaluation in terms of F1-score on the HIPE2022 dataset for Named Entity Recognition in English. Results are split into an earlier and a later period taking 1840 as the median date of the texts in the dataset.

We focused on the training and development splits from the topres19th subset. This subset consists of British Library newspapers from the 18^{th} and 19^{th} centuries, and the named entities correspond exclusively to geographical locations Ardanuy et al. [2022]. A total of approximately 3,300 entities are annotated, with 236 being reserved for development. We fine-tune the different models to perform token-level classification over sentences, with a total of 5,874 sentences for training and 646 for development. We fine-tune each model over 5 epochs and perform a total of 10 fine-tuning rounds per model in order to take into account random variation in the training procedure.

Figure 3 and Figure 4 show the results of this experiment. We report F1-score per model on the entire dataset (Figure 3) as well as the F1-score per model for 'earlier' and 'later' texts, taking the median year of the sentences as reference point (1840; see Figure 4). The difference in performance between the models is relatively small for NER. Overall, it appears TuringBERT is likely to outperform BERT-Adapted and possibly MacBERTh too. It should be noted, however, that the absolute difference in F1-scores is negligible. Furthermore, it is possible that TuringBERT's marginally superior performance is due to the fact that the NER data stems from the same collection as the pre-training data for TuringBERT. Thus, TuringBERT may be able to produce better features for entities that it has already processed in the pre-training phase. Finally, we observe that TuringBERT loses its marginal advantage when applied to earlier texts in the dataset (for which the results are overall worse for all models).

3.3 Word Sense Disambiguation

The next task we tackle is Word Sense Disambiguation (WSD), which we approach from a diachronic angle. For this – and the following – tasks, we rely on a custom evaluation dataset extracted from the Oxford English Dictionary [OED Simpson and Weiner, 1989]. With its large reservoir of sense distinctions, categorizations and exemplifications, the OED is an authoritative resource for historical and contemporary lexical semantics in English. We refer to Manjavacas and Fonteyn [2021] for the details on the compilation of the evaluation dataset used for the present experiments. In total, we evaluate the four candidate models on historical WSD from two distinct angles.

3.3.1 Non-Parametric Word Sense Disambiguation

The first historical WSD setting involves no fine-tuning, and relies entirely on vector similarity metrics to assign a target word in a given context to its corresponding sense. For a given historical input sentence like *"They must haue houses warme, as your Pigions haue, crossed through with small Pearches"*, exemplifying a sense of the word *"cross"*, we compute the contextualized vector representation of *"cross"* and measure its similarity to abstract vector representations of the different senses of the word *"cross"*. These abstract representations are computed as sense centroids, by averaging over the vector representations of the different exemplifications of a given sense in the OED dataset.

We evaluate the models using a total of 191 OED lemmata. These test lemmata contain all at least 50 example sentences and at least two different senses (the minimum number of senses required to perform word sense disambiguation). The resulting dataset consists of 17,878 sentences, which we split into a training and a test set, proportional to the number of sentences per sense. The training set is used to estimate the sense centroids. The test set is used to estimate the accuracy of this method for each of the language models.

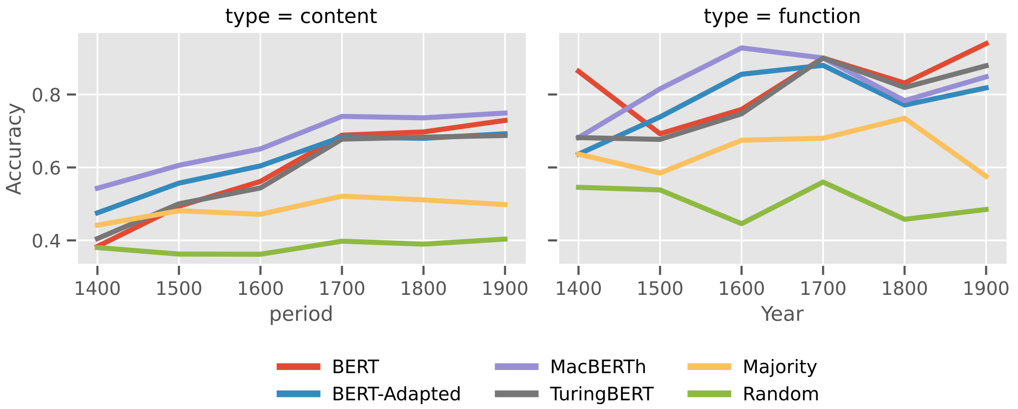

Figure 5: Results in terms of accuracy of the non-parametric WSD evaluation, differentiating between content and function words. On the x-axis, we aggregate over the time period of the corresponding test sentences. For comparison, a random and a majority baselines have been added.

Figure 5 shows the results of the experiment, differentiating between lemmata that belong to content words – i.e. nouns (e.g. *dog*, *sky*), adjectives (e.g. *nice*, *beautiful*) and verbs (e.g. *jump*, *think*) – and those that belong to function words – i.e. prepositions (e.g. *in*, *about*), pronouns (e.g. *me*, *who*), conjunctions (e.g. *and*, *since*) and interjections (e.g. *hey*). The reason why we differentiate between these two groups is that lexical or 'contentful' semantics can be considered distinct from grammatical or 'function' semantics, both in terms of how senses can be derived from contextual information (with the context of function words being more diverse than that of content words), and in terms of the diachronic dynamics [Hamilton et al., 2016].

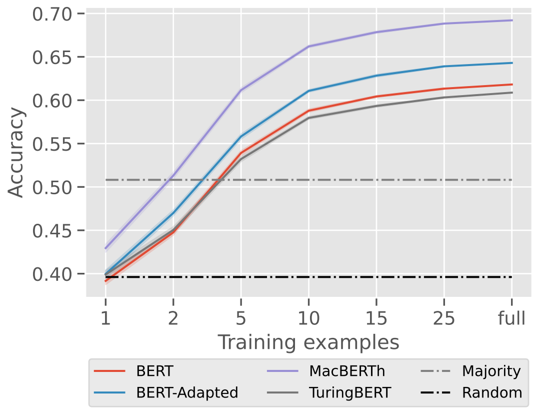

Figure 6: Results in terms of accuracy of the non-parametric WSD evaluation taking into account the number of sentences per sense in the training set. Majority and random baselines are added for reference.

Focusing on content words, we observe that all models perform above baseline level across all time periods, with MacBERTh outperforming the other models. Furthermore, BERT-Adapted has an advantage over the non-adapted variant BERT up until the 18^{th} century, where we observe a leap in performance with respect to the previous periods and a convergence of all models, indicating that the advantages of historical adaptation may be limited to the earlier periods. When considering function words, MacBERTh again seems to have an advantage – except in the earliest period and in the post-18^{th}century data. Beyond this, no clear patterns can be observed.

The effectiveness of this method is largely conditioned by the quantity and quality of the sense centroids. As we took an even split between training and test sets, we assumed that sense centroids can be estimated on data as large as the target data of interest. This is, however, unrealistic: in real-word scenarios, it is often the case that the amount of available labeled data – i.e. in this case, the number of example sentences per sense – is much smaller. In order to test whether models diverge in their requirements for training data, we ran an experiment in which we limit the number of sentences per sense to several values by sampling the target number of sentences 20 times. This gives us an indication of not only the amount of data required, but also of the dependency on the specific sentences in the training set for achieving strong performance.

Figure 6 shows the results of this experiment. For each target number of sentences on the x-axis, the y-axis shows the mean accuracy and its dispersion for each of the alternative models. Notably, all models are very robust against variation in the sentences used for estimating the centroids. As a result, there is very little variance in the obtained accuracy scores – i.e. there is almost no dispersion around the mean line –, even in the smaller data regimes. When examining the scores, we see that MacBERTh outperforms the other candidate models in all conditions, and is also able to outperform the strongest baseline – i.e. the majority baseline – with just 2 sentences per sense. After 10 sentences – i.e. approximately a fifth of the average number of sentences per sense in the dataset –, models start to converge to their optimal performance on this dataset. Again, BERT-Adapted shows a slight improvement over BERT, and TuringBERT performs in par with the non-historical model BERT.

3.3.2 Word-in-Context

The second approach to historical WSD reformulates the task as a binary task [Pilehvar and Camacho-Collados, 2019] (see also [Beelen et al., 2021], where the task is called "targeted sense disambiguation"). For any given pair of sentences exemplifying senses of the same lemma, we fine-tune our models to predict whether the two sentences exemplify the same word sense or not. In order to evaluate following this approach, we carve out a dataset from the OED similar to the one described by Manjavacas and Fonteyn [2021]. This dataset covers 416 lemmata and 84,712 sentences, from which we generate positive and negative pairs in the following manner: for each sentence in the dataset, we first sample a positive example from the set of sentences exemplifying the same sense. Subsequently, we sample a different sense belonging to the same lemma, and from the set of sentences illustrating that sense, we sample one negative example.

In order to fine-tune the models, we replicate the settings described in [Devlin et al., 2019, Section 4.1], using the last hidden activation corresponding to the [CLS] token, adding a linear projection layer in order to compute the probabilities that the sentences belong to the same sense, optimizing a cross entropy loss. In order to let the model focus on the word that corresponds to the underlying lemma, we add [TGT] tokens around the focus word in both members of the input pair.[6] This is exemplified by the sentences shown in Table 2.

The results of this experiment are shown in Figure 7, where we report accuracy numbers based on the centuries from which the left and right sentences stem (shown respectively on the y-axis and x-axis.) For each bin, we show the accuracy achieved by each model. Overall, the results are high, ranging from 85% to 98%. The plots highlight that the most difficult examples stem from the 15^{th} to 17^{th} centuries. In these bins, both MacBERTh and BERT-Adapted have an advantage.

In order to highlight the relative advantage of MacBERTh and BERT-Adapted on this task, Figure 8 and Figure 9 show, respectively, the differences in performance of MacBERTh and BERT-Adapted in comparison to the alternative models. Overall, MacBERTh outperforms the other models, with a larger performance difference in the earlier bins – i.e. the bottom-left part of the plot. The differences are surprisingly large when comparing with TuringBERT across all time periods. Moreover, BERT-Adapted has an advantage over BERT, especially when considering the earlier periods. This highlights the effectiveness of adapting present-day language

[6] We use the "sbert" library Reimers and Gurevych [2019] to fine-tune the models, training for 5 epochs with batch size of 16 on a single GPU.

	Left Quotation	Right Quotation
Example	He lov'd his Country with too unskilful a tenderness.	I love it to be grieved when he hideth his smiles.
Input	He [TGT] lov'd [TGT] his Country with too unskilful a tenderness.	I [TGT] love [TGT] it to be grieved when he hideth his smiles.
Sense	**1.a** *"To have or feel love towards (a person, a thing personified) (for a quality or attribute); to entertain a great affection, fondness, or regard for; to hold dear."*	**3.c** *"With direct object and infinitive or clause: to desire or like (something to be done). Also (chiefly U.S.) with for preceding the notional subject of the infinitive clause."*

Table 2: An example negative pair for lemma 'love' showcasing the modification in order to fine-tune the model.

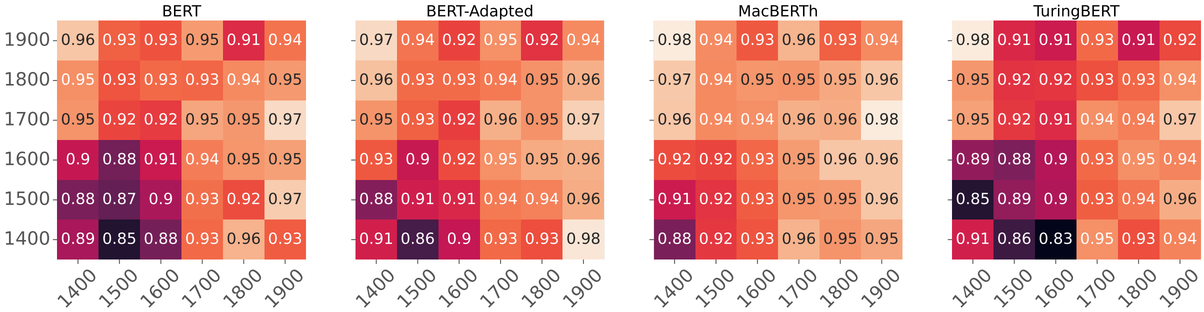

Figure 7: Accuracy in the Word-in-Context WSD task by period of the left and right examples shown respectively in the y-axis and the x-axis. The color matches the accuracy in the corresponding bins.

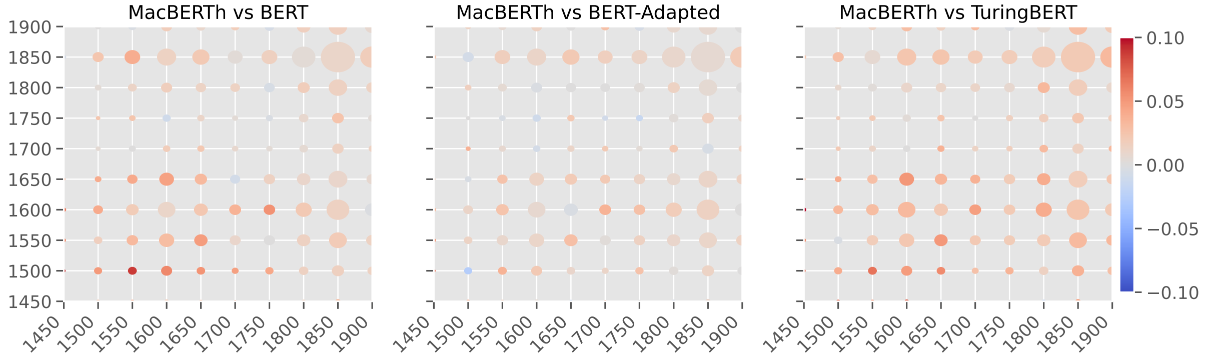

Figure 8: Circle plot showing the difference in accuracy between `MacBERTh` and the alternative models per period. The size of the circles correspond to the number of predictions in disagreement between the compared models. The color corresponds to the difference in accuracy.

models for targeted WSD. Still, the plots show that `MacBERTh` outperforms `BERT-Adapted` across all time periods, which offers an indication that pre-training from scratch is a stronger method than adaptation.

3.4 Fill-In-The-Blank

The next downstream task approaches Natural Language Understanding using a fill-in-the-blank evaluation scheme. This task does not require fine-tuning. Instead, we rely on the dataset of sense-exemplified quotations from the OED, and poll each model for the underlying lemma that the sentence is exemplifying after masking the word that corresponds to that lemma. The

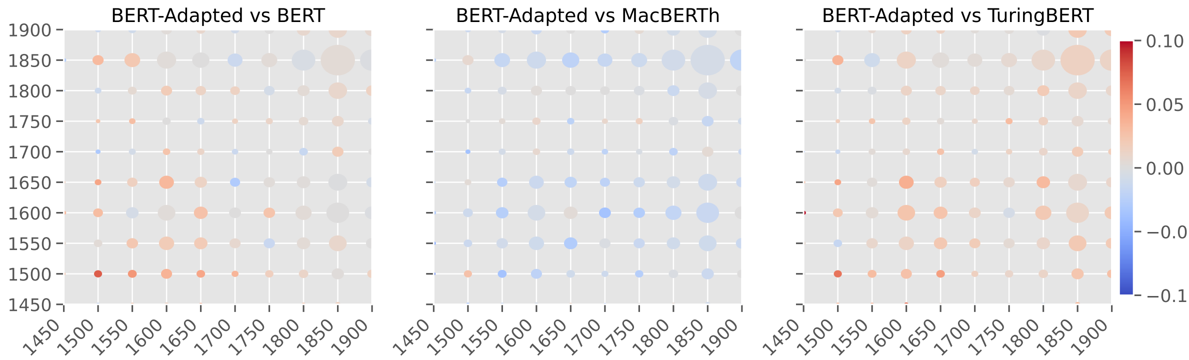

Figure 9: Circle plot showing the difference in accuracy between `BERT-Adapted` and the alternative models per period. The size of the circles correspond to the number of predictions in disagreement between the compared models. The color corresponds to the difference in accuracy.

models, thus, need to gather as much information as possible from the context in order to produce accurate guesses. Importantly, since the sentences are chosen in order to illustrate a particular usage of the given word, we can assume that they contain – otherwise the task would become artificially difficult, and less informative for benchmarking purposes.

Following the original masking loss [Devlin et al., 2019], a language model outputs a probability distribution over the model's own vocabulary for each masked word in the input. In order to assess the plausibility that each model assigns to the true target word, we compute its rank in this probability distribution – i.e. we do not use the probability itself, since this quantity would be difficult to compare due to differing vocabularies. For similar reasons, we restrict ourselves to sentences in which the target word is not sub-tokenized by any of the models. The resulting dataset comprises 42,961 sentences, covering 731 different words (with an average of 58.8 sentences per word.)

In order to summarize the model performance, we compute the Mean Reciprocal Rank (MRR), which in this case corresponds to averaging over the inverse of the individual ranks. Again, we factor the time dimension into the evaluation, binning the results into spans of 50 years. Figure 10 shows the results of this experiment. Here, `MacBERTh` has an advantage across all periods, except in the most recent bin – where all models seem to converge. `BERT-Adapted` also improves over the non-adapted variant (`BERT`), except for the 19^{th} century.

3.5 Sentence Periodization

Finally, we submit the models to a sentence periodization task. For a given input sentence, we fine-tune the models to make predictions about the year in which they were written. Algorithmically, we approach this task using a two-step setup.[7] First, we fine-tune the language models to perform a binary task in which the goal is to predict whether the first of two sentences stems from a later period than the second. Then, in order to predict the year of a given input sentence, we run the binary classifier comparing the input sentence with each of the sentences in a separate corpus – the background corpus. This background corpus has been sampled so as to have an even distribution of sentences over time periods, and consists of 5,000 sentences. Finally, we use the individual predictions comparing the input sentence to the sentences in afore-mentioned background corpus in order to construct a single year prediction over the entire range of years in the background corpus. The latter step uses the cumulative distribution of predictions and the

[7] Our first attempts involved using an ordinal regression on top of language model sentence embeddings, but results were not informative.

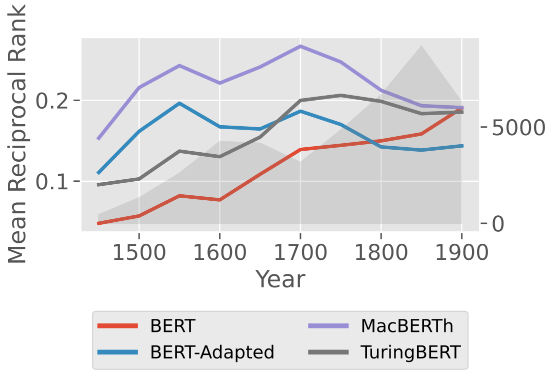

Figure 10: Results of the fill-in-the-blank task over time in terms of Mean Reciprocal Rank. Overlayed are the number of sentences included in each bin.

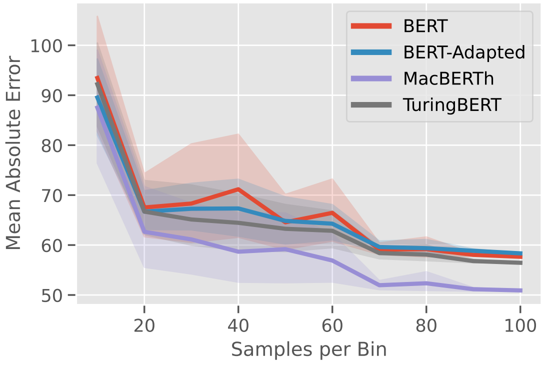

Figure 11: Mean Absolute Error (on the y-axis) on the sentence periodization task using background corpora of increasingly larger sizes (on the x-axis).

knee method for finding the cutoff point in this distribution [Satopaa et al., 2011] – we refer to the original paper for more details on this method [Manjavacas and Fonteyn, 2021].

We train the models using the cross-encoder implementation provided by “sbert” [Reimers and Gurevych, 2019]. The training data consists of 100,000 sentence pairs sampled from the OED dataset. The test set is constructed in a similar way, comprising a total of 5,000 sentence pairs.

The performance of this method depends on the size of the background corpus. In order to quantify this dependency, we ran a first experiment where we computed the Mean Absolute Error (MAE) over increasingly larger sub-samples of the background corpus (all sub-samples are uniformly sampled over the entire range to ensure that no time spans are over-represented.) For each size, we re-ran the experiment 10 times in order to quantify the dispersion due to the background corpus sample. Figure 11 shows the results of this experiment. We observe that all models reach their top performance when using the full background corpus – i.e. a total of 5,000 instances with 100 instances per each bin of 50 years – and very small dispersion is present starting with 70 instances per bin. Comparing the models on the full background corpus, we observe that `MacBERTh` has an advantage of slightly below a mean absolute error of 10 years.

Figure 12 shows finer-grained results of this experiment, factoring in the time dimension. We also include a baseline that predicts years randomly following the distribution of years in the test set. The random baseline reflects the fact that predicting sentences towards the middle of the range results in inherently smaller MAE. In order to remove this artifact from the visualization, Figure 13 reports the relative improvement of each model over the random baseline. This modification does not directly affect the comparison between the models, while letting us assess when the differences between the models are located in easier or more difficult periods.[8] Most of `MacBERTh`’s advantage is located in the years starting in 1750. Before that period, the differences between the models are small, with `MacBERTh` and `BERT-Adapted` at the top. Interestingly, the results for `BERT-Adapted` dip towards the later section of the background corpus.

[8] Note that in contrast to the original MAE scores, now higher means better – i.e. larger improvement.

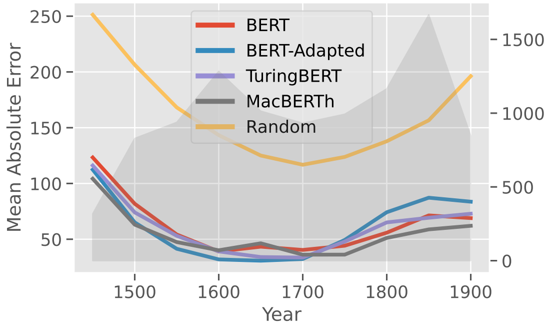

Figure 12: Mean Absolute Error of the compared models binned over periods of 50 years. We also include results for a random baseline.

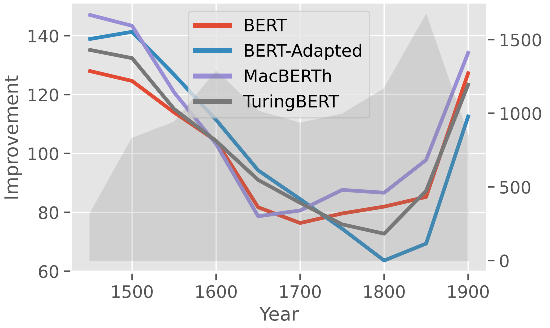

Figure 13: Average improvement in Mean Absolute Error of the different models over the random baseline.

IV DISCUSSION & CONCLUSION

Summarizing over the results of the various experiments, it seems reasonable to state that the most reliable means of making a BERT model suitable for applications to historical text is to pre-train a BERT model from scratch on historical corpus data. Overall, the historically pre-trained model MacBERTh had an advantage over the competitor models in POS-tagging, both types of Word Sense Disambiguation, Filling-In-the-Blank, and the Sentence Periodization tasks. The sole exception to these result is the NER evaluation, where TuringBERT performed marginally better on the post-1840 data. Yet, TuringBERT's advantage in NER may in fact be due to overlap in the evaluation data and the data used to adapt TuringBERT.

In certain cases – and particularly in POS-tagging – the advantages of MacBERTh were stronger when the training data available for fine-tuning was scarce. Since the pre-training dataset of MacBERTh covers a larger diachronic window than those of BERT and TuringBERT, it can be assumed that this dataset represents a more varied collection of texts, which could potentially explain the stronger sample efficiency that MacBERTh seems to have. Still, BERT-Adapted – a model adapted to the same pre-training dataset – does not seem to profit from this diversity as much, which may be due to imported biases derived from either the original present-day pre-training dataset or the tokenizer.

Interestingly, BERT-Adapted sometimes underperforms in the later periods – this is the case for the Word-in-Context (see Figure 9), Fill-in-the-Blank (see Figure 10) and the Sentence Periodization (see Figure 13) tasks. The cut-off point seems to be at around the 1800s.

While it is interesting in itself to state that there is a difference in the quality of the embeddings generated by the historically pre-trained model MacBERTh and the historically adapted models (TuringBERT), these results do raise the question of what exactly causes this difference. A similar result was obtained by the development team of the SciBERT model for Biomedical NLP [Beltagy et al., 2019]. This model was pre-trained from scratch on a large corpus of research papers mined from Semantic Scholar [Ammar et al., 2018] amounting to ca. 3.17B tokens. In this investigation, SciBERT was compared to BioBERT [Lee et al., 2020], another domain-specific model for Biomedical NLP that was adapted from BERT to an even larger corpus of articles from PubMed emcompassing ca. 18B tokens. Their results also highlight that the adapted model was subpar to the model that was pre-trained from scratch, even though the pre-training dataset of the later an order of magnitude smaller.

In this respect, a hypothesis that would need more thorough testing is whether the current tokenization approaches may hinder a successful adaptation of pre-trained model to different domains. Sun et al. [2020] showed that BERT models are brittle in the presence of misspelling, especially when the misspelling resulted in particularly awkward sub-word tokenization. In the case of historical material, Baptiste et al. [2021] have shown that `CharBERT` [Ma et al., 2020] – a BERT variant that processes input tokens character by character – produces more robust results against the presence of variation stemming from OCR noise. More generally, current research efforts have focused on producing tokenization-free models, which do not require language (or domain) specific tokenizers and can be thus applied more robustly across languages [Clark et al., 2022]. It remains to be tested whether these models have the potential to be efficiently adapted to new domains, and whether doing so produces more powerful features than models that are pre-trained from scratch.

Finally, it is also important to note that, while the historically adapted models were generally outperformed by `MacBERTh`, both `TuringBERT` and `BERT-Adapted` still showed some advantage over the non-adapted, present-day English model `BERT`. That historical adaptation (and adaptation more generally) is still a fruitful undertaking is of great value for the DH community: in some cases, adaptation is the only possibility, due to the sparsity of text in the target domain [e.g. Brandsen et al., 2021].

ACKNOWLEDGMENTS

The creation of MacBERTh has been made possible by the Platform Digital Infrastructure (Social Sciences and Humanities) fund (PDI-SSH). We want to thank the organizers and audience of the NLP4DH workshop, and the members of Text mining and Retrieval group (TMRL) from the University of Leiden for their valuable feedback on earlier versions of this paper. In particular, we thank Hugo de Vos for suggesting the addition of a random baseline in the Sentence Periodization task.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. DocBERT: BERT for Document Classification. *arXiv:1904.08398 [cs]*, August 2019. URL http://arxiv.org/abs/1904.08398. arXiv: 1904.08398.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-3011. URL https://aclanthology.org/N18-3011.

Mariona Coll Ardanuy, David Beavan, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, Katherine McDonough, Federico Nanni, Daniel van Strien, and Daniel CS Wilson. A dataset for toponym resolution in nineteenth-century english newspapers. *Journal of Open Humanities Data*, 8, 2022.

Blouin Baptiste, Benoit Favre, Jeremy Auguste, and Christian Henriot. Transferring modern named entity recognition to the historical domain: How to take the step? In *Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, 2021.

Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 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. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.243. URL https://aclanthology.org/2021.findings-acl.243.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José Moreno, Nicolas Sidère, and Antoine Doucet. Robust named entity recognition and linking on historical multilingual documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, volume 2696, pages 1–17. CEUR-WS Working Notes, 2020.

Alex Brandsen, Suzan Verberne, Karsten Lambers, and Milco Wansleeben. Can BERT Dig It? – Named Entity Recognition for Information Retrieval in the Archaeology Domain. *arXiv:2106.07742 [cs]*, June 2021. URL http://arxiv.org/abs/2106.07742. arXiv: 2106.07742.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Transactions of the Association for Computational Linguistics*, 10:73–91, 01 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00448. URL https://doi.org/10.1162/tacl_a_00448.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 288–310. Springer International Publishing, 2020a. ISBN 978-3-030-58219-7.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Overview of clef hipe 2020: Named entity recognition and linking on historical newspapers. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 288–310, Cham, 2020b. Springer International Publishing. ISBN 978-3-030-58219-7.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. Named entity recognition and classification on historical documents: A survey. *arXiv preprint arXiv:2109.11406*, 2021.

Maud Ehrmann, Matteo Romanello, Antoine Doucet, and Simon Clematide. *HIPE 2022 Shared Task Participation Guidelines*, February 2022. URL https://doi.org/10.5281/zenodo.6045662.

Anton Ehrmanntraut, Thora Hagen, Leonard Konle, and Fotis Jannidis. Type- and Token-based Word Embeddings in the Digital Humanities. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2021)*, volume 2989 of *CEUR Workshop Proceedings*, pages 16–38, 2021. URL http://ceur-ws.org/Vol-2989/long_paper35.pdf.

Katrin Erk. Vector Space Models of Word Meaning and Phrase Meaning: A Survey: Vector Space Models of Word and Phrase Meaning. *Language and Linguistics Compass*, 6(10):635–653, October 2012. ISSN 1749818X. doi: 10.1002/lnco.362. URL https://onlinelibrary.wiley.com/doi/10.1002/lnco.362.

Lauren Fonteyn. What about grammar? Using BERT embeddings to explore functional-semantic shifts of semi-lexical and grammatical constructions. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, volume 2723 of *CEUR Workshop Proceedings*, pages 257–268, 2020. URL http://ceur-ws.org/Vol-2723/short15.pdf.

Philip Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3960–3973, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.365. URL https://www.aclweb.org/anthology/2020.acl-main.365.

Edouard Grave. Language Identification · fastText. https://fasttext.cc/blog/2017/10/02/blog-post.html, 2017.

Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. Novel2vec: Characterising 19th century fiction via word embeddings. In *24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), University College Dublin, Dublin, Ireland, 20-21 September 2016*, 2016.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2116–2121, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1229. URL https://www.aclweb.org/anthology/D16-1229.

Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. 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 4238–4248, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1433. URL https://aclanthology.org/D19-1433.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. Neural Language Models for Nineteenth-Century English. *Journal of Open Humanities Data*, 7:22, September 2021a. ISSN 2059-481X. doi: 10.5334/johd.48. URL http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.48/.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. Neural Language Models for Nineteenth-Century English (dataset; language model zoo). https://doi.org/10.5281/zenodo.4782245, May 2021b.

Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C Dubnicek, Ted Underwood, and J Stephen Downie. Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2021)*, volume 2989 of *CEUR Workshop Proceedings*, pages 266–279, 2021. URL http://ceur-ws.org/Vol-2989/long_paper43.pdf.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Leonard Konle and Fotis Jannidis. Domain and Task Adaptive Pretraining for Language Models. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, volume 2723 of *CEUR Workshop Proceedings*, pages 248–256, 2020. URL http://ceur-ws.org/Vol-2723/short33.pdf.

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. Penn-helsinki parsed corpus of early modern english, 2004.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*, 2018.

Kai Labusch, Clemens Neudecker, and David Zellhöfer. Bert for named entity recognition in contemporary and historical german. In *Proceedings of the 15th Conference on Natural Language Processing, Erlangen, Germany*, pages 8–11, 2019.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240, 2020.

Alessandro Lenci. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4(1):151–171, 2018. doi: 10.1146/annurev-linguistics-030514-125254. URL https://doi.org/10.1146/annurev-linguistics-030514-125254. _eprint: https://doi.org/10.1146/annurev-linguistics-030514-125254.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. Charbert: Character-aware pre-trained language model. *arXiv preprint arXiv:2011.01513*, 2020.

Enrique Manjavacas and Lauren Fonteyn. Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950). In *Proceedings of the Workshop on NLP4DH @ ICON 2021*, online, December 2021. NLP Association of India (NLPAI).

Jani Marjanen, Lidia Pivovarova, Elaine Zosa, and Jussi Kurunmaki. Clustering Ideological Terms in Historical Newspaper Data with Diachronic Word Embeddings. In *The 5th International Workshop on Computational History (HistoInformatics 2019)*, volume 2461 of *CEUR Workshop Proceedings*, pages 21–29, 2019. URL http://ceur-ws.org/Vol-2461/paper_4.pdf.

Carlos Martinez-Ortiz, Tom Kenter, Melvin Wevers, Pim Huijnen, and Joris van Eijnatten. Design and implementation of ShiCo: Visualising shifting concepts over time. In *The 5th International Workshop on Computational History (HistoInformatics 2019)*, volume 1632 of *CEUR Workshop Proceedings*, pages 11–19, 2019. URL http://ceur-ws.org/Vol-1632/paper_2.pdf.

Janis Pagel, Nidhi Sihag, and Nils Reiter. Predicting Structural Elements in German Drama. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2021)*, volume 2989 of *CEUR Workshop Proceedings*, pages 217–227, 2021. URL http://ceur-ws.org/Vol-2989/short_paper34.pdf.

Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1128. URL https://aclanthology.org/N19-1128.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. Tracing semantic change with Latent Semantic Analysis. In Kathryn Allan and Justyna A. Robinson, editors, *Current Methods in Historical Semantics*. DE GRUYTER, Berlin, Boston, January 2011. ISBN 978-3-11-025290-3. doi: 10.1515/9783110252903.161.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171, June 2011. doi: 10.1109/ICDCSW.2011.20.

Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

Stefan Schweter and Johannes Baiter. Towards robust named entity recognition for historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4312. URL https://www.aclweb.org/anthology/W19-4312.

Stefan Schweter and Luisa März. Triple e-effective ensembling of embeddings and language models for ner of historical german. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, volume 2696. CEUR-WS Working Notes, 2020.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

John Simpson and Edmund Weiner. *Oxford English Dictionary*. Oxford University Press, 1989.

Matthew Sims, Jong Ho Park, and David Bamman. Literary Event Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1353. URL https://www.aclweb.org/anthology/P19-1353.

Pia Sommerauer and Antske Fokkens. Conceptual Change and Distributional Semantic Models: An Exploratory Study on Pitfalls and Possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-4728. URL https://www.aclweb.org/anthology/W19-4728.

Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *arXiv preprint arXiv:2003.04985*, 2020.

Nina Tahmasebi and Thomas Risse. On the Uses of Word Sense Change for Research in the Digital Humanities. In Jaap Kamps, Giannis Tsakonas, Yannis Manolopoulos, Lazaros Iliadis, and Ioannis Karydis, editors, *Research and Advanced Technology for Digital Libraries*, volume 10450, pages 246–257. Springer International Publishing, Cham, 2017. ISBN 978-3-319-67007-2 978-3-319-67008-9. doi: 10.1007/978-3-319-67008-9_20. URL http://link.springer.com/10.1007/978-3-319-67008-9_20. Series Title: Lecture Notes in Computer Science.

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

P. D. Turney and P. Pantel. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188, February 2010. ISSN 1076-9757. doi: 10.1613/jair.2934. URL https://jair.org/index.php/jair/article/view/10640.

Joris van Eijnatten and Ruben Ros. The Eurocentric Fallacy. A Digital-Historical Approach to the Concepts of 'Modernity', 'Civilization' and 'Europe' (1840–1990). *International Journal for History, Culture and Modernity*, 7(1):686–736, November 2019. ISSN 2666-6529, 2213-0624. doi: 10.18352/hcm.580. URL https://brill.com/view/journals/hcm/7/1/article-p686_33.xml.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Melvin Wevers. Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. *arXiv:1907.08922 [cs, stat]*, July 2019. URL http://arxiv.org/abs/1907.08922. arXiv: 1907.08922.

Melvin Wevers and Marijn Koolen. Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(4):226–243, October 2020. ISSN 0161-5440, 1940-1906. doi: 10.1080/01615440.2020.1760157. URL https://www.tandfonline.com/doi/full/10.1080/01615440.2020.1760157.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In

Proceedings of the IEEE international conference on computer vision, pages 19–27, 2015.