MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950)

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

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

Abstract

The new pre-train-then-fine-tune paradigm in Natural Language Processing (NLP) has made important performance gains accessible to a wider audience. Once pre-trained, deploying a large language model presents comparatively small infrastructure requirements, and offers robust performance in many NLP tasks. The Digital Humanities (DH) community has been an early adapter of this paradigm. Yet, a large part of this community is concerned with the application of NLP algorithms to historical texts, for which large models pre-trained on contemporary text may not provide optimal results. In the present paper, we present "MacBERTh"—a transformer-based language model pre-trained on historical English—and exhaustively assess its benefits on a large set of relevant downstream tasks. Our experiments highlight that, despite some differences across target time periods, pre-training on historical language from scratch outperforms models pre-trained on present-day language and later adapted to historical language.[1]

1 Introduction & Related Work

Social scientists and Humanities scholars have long been interested in describing cultural systems and understanding the way in which these change across time. Traditionally, such shifts were documented with 'manual' interpretative methods, but more recently researchers in DH have begun applying Machine Learning techniques to support their interpretation.

In the case of researchers working with historical text, current work has been occupied with developing and evaluating NLP algorithms with the goal of modeling the way in which concepts, categories and discourses (e.g. of class, gender) change over time along with their linguistic representations.

In this context, applications include data-driven approaches to conceptual change (Fitzmaurice et al., 2017; Sommerauer and Fokkens, 2019; Marjanen et al., 2019; Martinez-Ortiz et al., 2019), historical word sense disambiguation (Bamman and Crane, 2011; Fonteyn, 2020; Beelen et al., 2021), Named-Entity Recognition in historical text (Labusch et al., 2019; Konle and Jannidis, 2020; Schweter and Baiter, 2019; Schweter and März, 2020; Ehrmann et al., 2020; Boros et al., 2020) or unsupervised semantic change (Schlechtweg et al., 2020; Giulianelli et al., 2020).

In view of the growing weight of the new NLP paradigm of "pre-train-and-fine-tune"—which leverages large language models in order to produce strong feature extractors (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019)—, the question arises as to whether similar performance boosts can be gained for current NLP-oriented research dealing with historical text.

Due to the heavy domain-shift along grammatical, semantic and orthographic language layers, models pre-trained on contemporary data are less helpful when applied on historical data. Previous work has experimented with adapting contemporary models to historical data on a per-task basis (Han and Eisenstein, 2019), although it is unclear whether this approach can yield a general purpose historical model. An alternative approach would be to adapt historical text to the modern standards using a historical text normalization system (Bollmann, 2019). While this could indeed help tackling orthographic shifts, grammatical and semantic shifts would be left un-adapted. Moreover, error percolation from the automatic normalization system would still be an issue.

Arguably, the main advantage of pre-training

[1] Evaluation code is available through the project's repository: https://www.github.com/emanjavacas/macberth-eval. "MacBERTh" itself is available as emanjavacas/MacBERTh from the transformers repository (Wolf et al., 2019).

Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 23–36
co-located with ICON-2021, December 16-19, 2021, Silchar, India.
https://doi.org/10.26615/978-952-94-5833-2_004

this type of models is their ability to exploit very large datasets. In the case of historical linguistic resources, the known scarcity of digitized text—even for commonly high-resource languages like English—turns such enterprise problematic. However, ongoing efforts towards making historically relevant book collections digitally accessible (Langley and Bloomberg, 2007; Labs, 2014; Mueller et al., 2016) have kick-started experimentation in this respect. For example, Bamman and Burns (2020) pre-trained a model on Latin text spanning several centuries. Schweter and Baiter (2019) and Konle and Jannidis (2020) have employed contextualized character-level models, (Schweter, 2020) has released historical German and French models trained on historical newspaper data, and Beelen et al. (2021) and Hosseini et al. (2021a) trained and released models on an English corpus spanning the 18th to 20th centuries.

Several questions appear in this context. For example, it remains unclear whether all target periods can benefit equally from a "historically" pre-trained model or whether the performance benefits of these models vary across periods depending on the available amount and type of documents. Moreover, it is unclear what the advantages are of the two current alternative pre-training approaches. In some cases, pre-existing models pre-trained on contemporary datasets are first "historically fine-tuned" before being applied on downstream tasks. This approach—motivated by the promise to leverage a larger out-of-domain contemporary dataset—has been shown to outperform their non-adapted counterparts (although, see German BERT vs. Europeana BERT in Schweter and März, 2020), but it remains unclear how these fine-tuned models compare to models pre-trained "historically" from scratch, and, more generally, whether the presence of modern data in the training process diminishes model performance (as suggested by Boros et al., 2020).

Contributions In this paper, we introduce a model pre-trained on a large span of historical English (1450-1900), and show its advantages with respect to present-day models, as well as models adapted from present-day to historical English on an exhaustive set of ad-hoc downstream tasks. Moreover, we show how model performance strongly depends on the time period of the target application.

2 Experimental Setup

We rely on the large language model known as "BERT"—a stack of transformer layers with a self-attention mechanism (Vaswani et al., 2017), optimizing a Masked Language Model (MLM) objective (Devlin et al., 2019). Despite the existence of several MLM alternatives, BERT remains a good choice, considering that (i) it is well-established and most thoroughly studied, and (ii) on-going evaluation of alternative choices—mostly focus on Natural Language Understanding (NLU) tasks—has not yielded a clearly superior architecture.

We rely on the seminal implementation,[2] with the hyper-parameterization corresponding to the "BERT-base Uncased" architecture.[3] Pre-training is done with default parameters, except for the maximum sequence length (set to 128 subtokens) for 1,000,000 training steps.

2.1 Pre-training Dataset

The model is pre-trained on a corpus of a total size of ca. 3.9B (tokenized) words (time span: 1450-1950) using the following corpora: the Early English Books Online (EEBO) corpus, the Corpus of Late Modern English Texts (CLMET3.1), the Evans Early American Imprints Collection (EVANS), Eighteenth Century Collections Online (ECCO), the Corpus of Historical American English (COHA), and the Hansard corpus (Hansard). The resulting corpus is a varied collection in terms of text types, including literary works, religious and legal text as well as news reports and transcriptions of British parliamentary debates. The summary word count statistics are shown in Figure 1.

In terms of preprocessing, the corpus was first cleaned up in order to remove foreign text,[4] and split into sentences using the NLTK built-in sentence tokenizer (Bird, 2006).

2.2 Benchmarking

To cast light upon the advantages of large MLMs for diachronic tasks, we designed a number of benchmarking tasks, in which the contextualized

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

[3] See Section 2.2.2 or the original paper for a description of these parameters.

[4] 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 language.

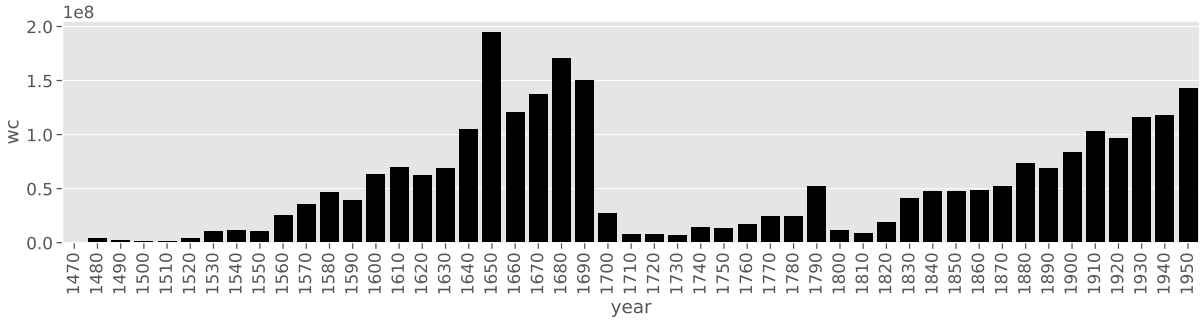

Figure 1: Aggregated word count statistics per decade for the pre-training corpora used in the present study.

representations of candidate models must encode historical information in order to achieve strong performance.

2.2.1 Benchmarking Datasets

In line with previous work (Hu et al., 2019; Heo et al., 2020; Beelen et al., 2021), we rely on data from the Oxford English Dictionary (OED Simpson and Weiner, 1989)—an authoritative resource for historical and contemporary lexical semantics in the English lexicon—for the benchmarking tasks. For each lemma, the OED defines a hierarchy of word senses, including quotations exemplifying each sense over the entire historical span of that sense. For the present experiments, we sampled 3,000 words from the vocabulary of the corpus described in Section 2.1, in proportion to their smoothed relative frequencies. Each word was retrieved and matched to existing lemmata in the OED reservoir. Upon successful retrieval, the senses and quotations of the corresponding lemma were stored. The resulting dataset comprises 2,700 lemmas, 35,110 senses and 246,048 quotations, which we utilize in varied ways for benchmarking.

We also include part-of-speech tagging, using the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME) (Kroch et al., 2004)—a manually annotated corpus of Early Modern English (time span: 1450-1700), comprising about 1.7m words over 448 documents.[5]

2.2.2 Candidate Models

In order to quantify the relative advantage of pre-training MLMs on historical corpora, we compare different instantiations of BERT. First, we benchmark against models with comparable architectures trained on **present-day English data** only. We use `BERT`, which corresponds to "BERT-Base Uncased" in the original repository, and 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; and `MultiBERT`, which corresponds to "BERT-Base Multilingual Cased" from the original repository, and is trained on the union of the top 100 languages in terms of the size of the respective Wikipedia sites, using a shared WordPiece vocabulary of 110,000.

Secondly, we compare with a variant of `BERT`—i.e. "BERT-Base Uncased"—that was fine-tuned at the Alan Turing Institute on 5.1B tokens of **historical English** (time span: 1760-1900 Hosseini et al., 2021a),[6] which we label `TuringBERT`.

Contemporary `BERT` and `MultiBERT` differ mainly in training material and vocabulary. `MultiBERT` should have an advantage when applied to historical English data, as it is trained on a much larger and varied dataset and with a more flexible vocabulary. The main difference between `TuringBERT` and `MacBERTh` is their span and size, with `TuringBERT` covering a smaller time window but a larger training dataset. Furthermore, as `MacBERTh` was pre-trained from scratch, its vocabulary is better adjusted to historical English.

3 Results

We now describe the benchmarking tasks and the results of the competing models in detail.

3.1 Part-of-speech Tagging

The first task tackles part-of-speech tagging of historical documents. Historical text is known to be challenging for automatic processing due

[5]We replicate the training-test splits from (Han and Eisenstein, 2019), thus keeping 115 files for testing and reserving from the 333 remaining 17 randomly sampled files (ca. 5%) for development purposes.

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

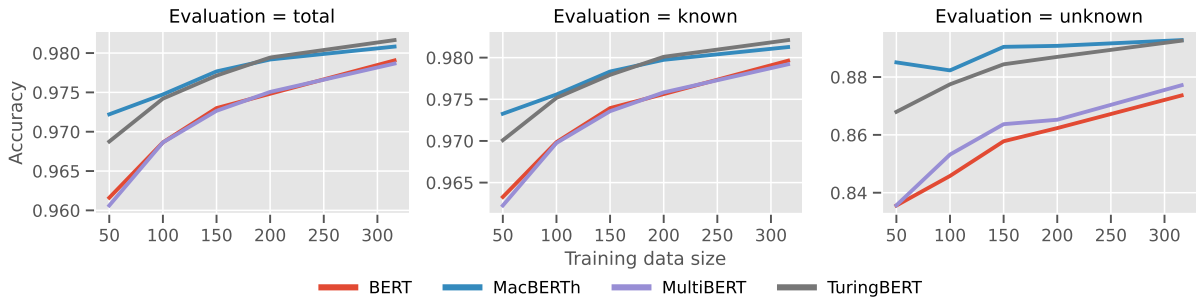

Figure 2: Accuracy on the part-of-speech tagging task (on the y-axis) considering different training size regimes (in number of documents in the x-axis) and different evaluation criteria: total (all tokens), known (only tokens observed in the training set) and unknown (only tokens not present in the training set).

to (relatively more) complex inflection systems and lacking orthographic standards (Manjavacas et al., 2019). As recently shown by Han and Eisenstein (2019), domain-specific pre-training yields improvements for historical pos-tagging, even if no labeled data is available for the target domain. We quantify the potential of pre-training on historical data for pos-tagging of historical texts by computing accuracy on held-out data after fine-tuning the pre-trained MLMs on **target-domain** data.[7]

We expect pre-training on target domain data to improve tagging accuracy of documents, especially if these stem from the same period. Moreover, in line with (Han and Eisenstein, 2019) we particularly expect improvements for tokens in the held-out data that were not encountered during training. Finally, we also test the relative sample efficiency of the competing MLMs by fine-tuning these on incrementally smaller samples of training data. To test our hypotheses, we compute (micro-)accuracy of all, known and unknown tokens, using random sub-samples (50, 100, 150, 200 and all files).[8]

Pre-training on contemporary material (`BERT`, `MultiBERT`) results in less accurate models for all evaluation conditions (see Figure 2). The historical models have an advantage, especially on unknown tokens. The model pre-trained on the larger temporal span (i.e. `MacBERTh`) has an advantage in the smaller training data regime.

Moreover, when factoring in the date of the held-out document, we find that the relative improvement of `MacBERTh` is larger for earlier dates, and seems to increase for later dates, as shown in Figure 3 for the accuracy of unknown tokens in the small-size training regime. This may be explained by the combined effect of the sample efficiency of the different models and the training data size of the different periods (shown in Figure 3 by the shaded grey area), which is seemingly correlated with the advantage of `MacBERTh`.[9]

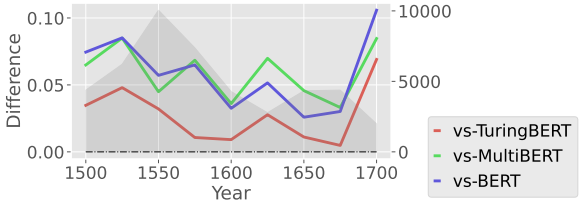

Figure 3: Difference in part-of-speech tagging accuracy of unknown tokens of `MacBERTh` with respect to the alternative models in the small-size training regime (50 documents). The shaded area shows the total number of tokens per period on which the evaluation is based.

3.2 Word Sense Disambiguation

Word Sense Disambiguation (WSD) by modeling lexical semantics in context has received ample attention from the DH community—see the different applications surveyed by Tahmasebi et al. (2018, Section 7.1 and 7.2)—, and is arguably one of the most promising venues for deploying MLMs.

We first approach WSD as a binary classification task in which pre-trained models are fine-tuned in order to predict whether a pair of quotations exemplifying senses of a given lemma correspond to the same sense or not (Section 3.2.1). Second, we evaluate the quality of sense embeddings derived from the MLMs without explicit fine-tuning in a sense classification task (Section 3.2.2).

[7] We fine-tune all MLMs on the PPCEME for 3 epochs with a batch size of 8 on a single GPU.

[8] The random sub-samples of training data were kept constant for all models.

[9] For the sake of completeness, the full results are shown in Figure 10 in the Appendix.

3.2.1 Word-in-Context

The word-in-context task for evaluating context-sensitive word representations was introduced by Pilehvar and Camacho-Collados (2019), following earlier efforts on evaluating context-dependent word similarity (Huang et al., 2012). Recently, Beelen et al. (2021) have also focused on this task, referring to it as "targeted sense disambiguation".

We utilize the OED quotations dataset from Section 2.2.1 in the following manner. First, we drop quotations with less than 5 words, in order to ensure that there is enough context for disambiguating. Second, we drop lemmata with less than 100 quotations left as well as lemmata that do not correspond to nouns, adjectives and verbs (based on the OED's lemma categorization). From the resulting dataset of 408 lemmata we reserved 10% (= 41), which are used for testing the generalization capabilities of the models. For a given input quotation, we generate a positive example by sampling a paired quotation belonging to the same sense and a negative example by sampling a quotation from a different sense of the same lemma.

In order to fine-tune the models, we replicate the settings 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 logits of the positive and negative class, and 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.[10,11]

Table 1 shows the results of development (= 25% of the training data) and held-out data. For each block of results, we further distinguish whether the instantiation of the lemma corresponds to the same part-of-speech tag, and, in case of correspondence, we report results per part-of-speech tag—i.e. noun (N), adjective (Adj) or verb (V).

Overall, `MacBERTh` obtained the best results across conditions, except for held-out adjectives where contemporary models had an advantage. Performance on development data is generally very high, surpassing 90% accuracy across conditions. However, on held-out lemmata, no model surpasses 70% accuracy (with adjectives being easier to classify than nouns and verbs for all models).

Model	Total	≠ POS	= POS		
			N	Adj	V
Development					
`BERT`	89.9	92.3	90.1	92.6	86.7
`MultiBERT`	92.0	94.8	91.6	95.6	88.7
`TuringBERT`	91.0	94.1	90.5	94.3	87.6
`MacBERTh`	**94.5**	**96.1**	**94.1**	**96.8**	**92.6**
Held-out					
`BERT`	59.5	56.8	60.5	65.8	58.4
`MultiBERT`	62.1	63.3	64.0	**66.1**	57.1
`TuringBERT`	58.6	58.5	59.8	60.7	56.4
`MacBERTh`	**63.0**	**63.8**	**65.3**	61.8	**59.3**

Table 1: Results of the word-in-context task for development and held-out lemmata across different conditions.

Pairs with diverging part-of-speech tags resulted in higher accuracy, which can be explained by class imbalance: pairs with diverging tags tend to belong to different senses. Interestingly, the drop in performance for the positive class with respect to the negative class was much smaller for `MacBERTh` (6.9 points) than for the other models (9.8 for `BERT`, 9.9 for `MultiBERT` and for 12 points for `TuringBERT`), thus suggesting a stronger generalization ability of `MacBERTh` over competitors.

Finally, we observe that the afore-mentioned advantage is not evenly distributed over the periods from which the input quotations stem. Instead, the advantage of `MacBERTh` was generally larger for quotations originating before the 1700s.[12]

3.2.2 Parameter-free WSD

In the full-fledged WSD setting, an input quotation must be tagged with the sense that it is exemplifying. A further difference with the word-in-context task is that we do not use any additional fine-tuning in order when approaching the task. Instead, we follow the approach outlined in Peters et al. (2018, Section 5.3). The distributed representations of senses are first computed, and then a sense is predicted for an unseen input quotation based on its proximity to the different sense representations (using the nearest sense representation neighbor in terms of cosine similarity). We restrict ourselves to a centroid approach to building sense representations, in which the contextualized vectors of the

[10] An example input pair can be seen in Table 3 in the Appendix.

[11] 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.

[12] For completeness, a full visualization of the difference in accuracy over time bins can be seen in Figure 11 (Appendix).

Model	Total	Word Type	
		Content	Function
BERT	36.0	36.1	34.8
MacBERTh	**42.3**	**42.0**	**50.4**
MultiBERT	32.3	32.1	38.6
TuringBERT	34.8	34.6	43.2
Majority	13.6	13.7	9.1
Random	9.2	9.3	6.1

Table 2: Results of the WSD comparison in terms of classification accuracy by word type.

target tokens exemplifying a particular word sense are averaged.

In order to build a dataset, we utilize the OED quotations from Section 2.2.1. We first drop lemmata with less than 50 quotations. Second, we discard single-sense lemmata as well as senses (of a given lemma) with less than 2 quotations (as we cannot produce classifications in those cases). On the basis of the remaining senses, we generate a stratified training and test set split with 50% of the quotations in each set. In order to classify the sense of an input sentence, we only need to compare it against the sense representations of the same lemma. For this purpose, we rely on the original OED's lemmata, thus assuming gold lemmata.

The results, split by word type, are shown in Table 2.[13] MacBERTh outperforms the competitors across all conditions. Overall, models performed better on function words than on content words, even though the latter seems to be an easier task as per the baseline. Interestingly, TuringBERT is outperformed by BERT, despite the former being fine-tuned on historical material.

Figure 4 factors in time on the x-axis, showing that the effect of time on accuracy is constant across models for content words. In the case of function words, the historical pre-training of MacBERTh seems to be of benefit in the earlier periods.

3.3 Fill-in-the-blank

The ad-hoc fill-in-the-blank task indirectly tackles NLU. For a given OED input, we mask the target token (i.e. the token corresponding to the word of which a sense is being exemplified) and interpret the plausibility assigned by the model to the target token as a proxy of the model's strength to capture

[13] We consider function words those tagged with "pron", "prep", "conj" or "int" following OED's classification.

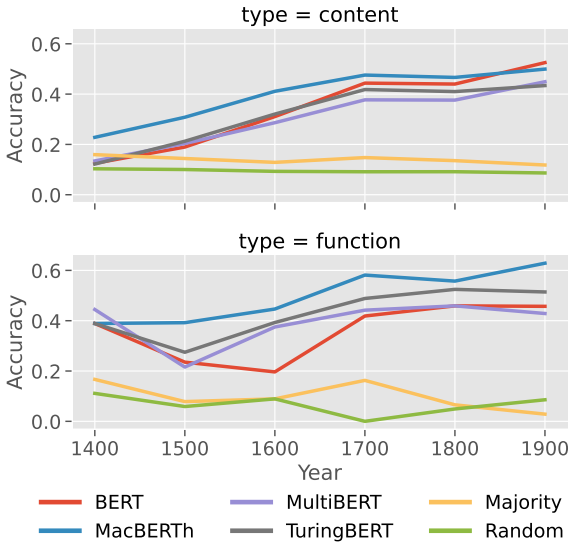

Figure 4: Results of the parameter-free WSD classification experiment by word type and year, including results for Majority and Random baselines.

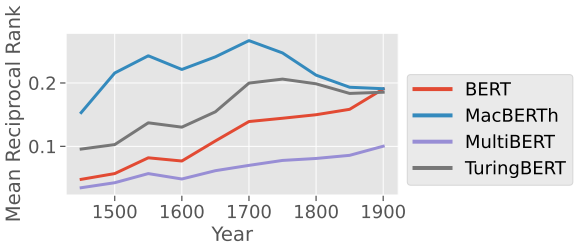

Figure 5: Results of the fill-in-the-blank task over time in terms of Mean Reciprocal Rank.

meaning.

Using the OED quotations from Section 2.2.1, we first select quotations for which the target token is part of the vocabulary of all compared models, which ensures that the comparison is fair. Moreover, since models differ in vocabulary sizes, we evaluate using the rank of the target token based on the logits (instead of directly using the full output distribution of logits). We use the Mean Reciprocal Rank as evaluation metric, averaging over quotations—shown in Figure 5.

The MacBERTh model tops across all periods. The difference with respect to the other models is larger in the earlier periods, highlighting a stronger ability to capture the semantics of earlier examples.

3.4 Sentence Periodization

The last task concerns periodizing quotations from the OED. OED quotations constitute a particularly well-suited test bed, as they have been selected by the OED editors in order to exemplify particular

word usages within specific diachronic frames. A periodizing model, thus, may exploit not only formal aspects of how English has changed (such as spelling, morphology, word order or grammar) but also changes in lexical semantics.

To tackle this task, we first fine-tune the different MLMs in a binary classification task with the goal of predicting whether the first sentence stems from a later period than the second. We deploy the same architecture as the one described in Section 3.2.1 but drop the signalization of the target token.

In order to periodize an input sentence, we use a subset of sentences for which the dates are known (we refer to this subset as the "background corpus"). This subset is both representative of the entire time range for which predictions need to be produced (i.e. sampled uniformly over equally sized spans in the OED), and kept apart during training. Then, for a given input sentence, we obtain a distribution of scores (i.e. probability) over years by comparing the input sentence against sentences from this background corpus.

For each background sentence, the model yields a probability that the input sentence stems from a later period. We first sort these probabilities by the years corresponding to the sentences in the background corpus. We then compute the cumulative distribution (which draws a strictly increasing curve). The prediction then corresponds to the point of maximum curvature or **knee point** within this curve, which we compute using the Kneedle method described by Satopaa et al. (2011). This method identifies the highest point in the curve after (i) smoothing out edges using a polynomial fit of the input data points and (ii) rotating the curve so that both the start and end point lie on the same horizontal line. An example prediction using this method is shown in Figure 6.

We use the OED data from Section 2.2.1, removing quotations with less than 5 words. From the remaining set, we reserve 5% for development and 5% for testing. The remaining 90%, is randomly split into 75% for training and 25% for the background corpus (which is produced by binning the range from 1450 to 1900 into decades and sampling 20 quotations per decade, giving a background corpus of 1,000 quotations in total). Finally, the training, test and development splits are turned into datasets by generating random pairs, ensuring that quotations in the input pairs do not belong to the same lemma. We restrict ourselves to 100,000 training and 5,000 development and test input pairs.[14]

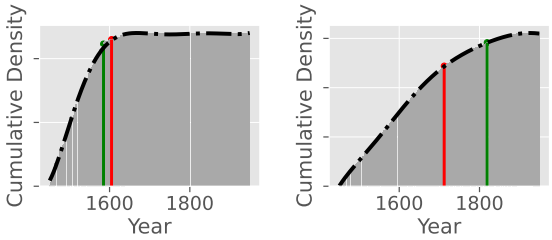

Figure 6: Visualization of a sentence periodization prediction using the knee method. The dashed line shows the cumulative distribution of prediction scores of a given input sentence with respect to the background corpus (x-axis). The grey line corresponds to the smoothing derived from a 7^{th} degree polynomial fit. Finally, the green and red lines highlight the true and predicted year, respectively. Left and right plots show examples of an accurate and inaccurate prediction.

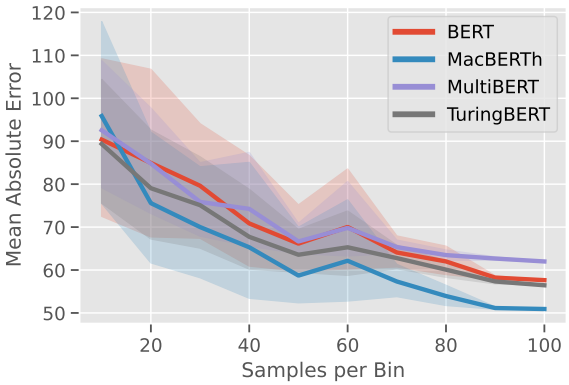

Figure 7: Visualization of the Mean Absolute Error achieved by the different models (lower is better) with respect to the number of samples in the background corpus.

Figure 7 shows the results in terms of Mean Absolute Error (MAE). As the size of the background corpus is a source of variation, we re-run the experiment varying the number of background instances within each 50 year bin (shown on the x-axis), until reaching the full size of 1,000 background quotations (i.e. 100 instances for each of the 10 bins). All models converge to their optimum performance when using the full background corpus. Figure 7 also shows that `MacBERTh` has the smallest error, being wrong on average by 50 years. `TuringBERT` is on par with `BERT` and outperforms `MultiBERT`.

Figure 9 factors in the time dimension, aggre-

[14] We fine-tune the models following the same setting as in Section 3.2.1.

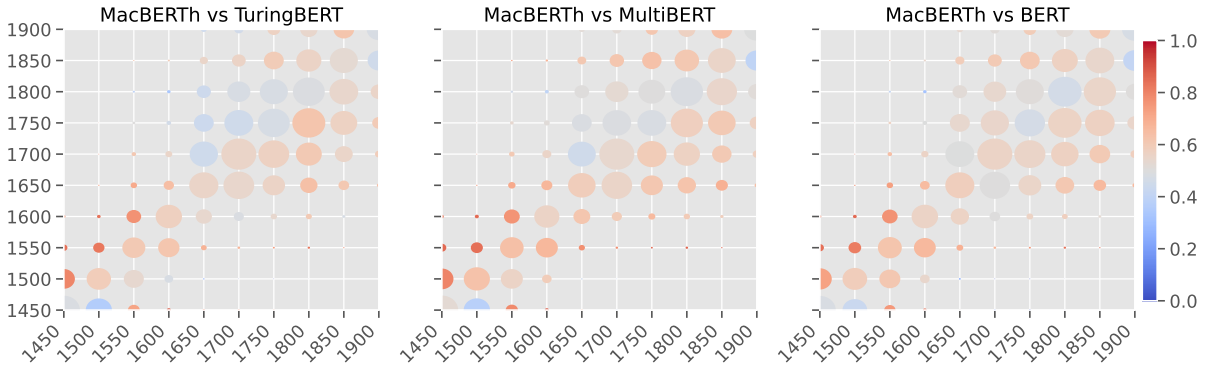

Figure 8: Visualization of the difference in performance of the competing models with respect to MacBERTh. Each circle corresponds to predictions in which the models being compared diverge. The color indicates the proportion of diverging predictions in which MacBERTh is right (thus, red and blue indicates whether MacBERTh is predominantly right). The size of each circle is proportional to the total number of diverging predictions. The y-axis and x-axis indicate respectively the period of the left and right input sentences.

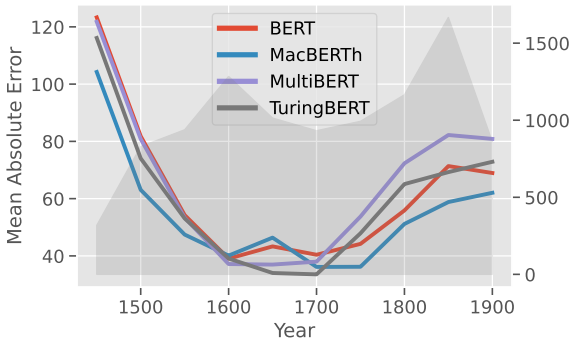

Figure 9: Mean Absolute Error over bins of 50 years. The shaded area shows the total number of pairs per period on which the evaluation is based.

gating MAE over bins of 50 years. MacBERTh achieves the best results for input sentences dated before 1650, as well for those dated after 1750. Further, Figure 8 shows the comparison of MacBERTh vs. the rest taking into account the source year bin of the left (y-axis) and right (x-axis) input examples. All models agree on the periodization of sentences that are separated in time by a larger span—i.e. off-diagonal bins are sparsely populated, indicating a small number of divergent predictions—especially when sentences prior to 1650 enter the comparison. MacBERTh's improvement over the other models seems to be concentrated in the period before 1650 (as suggested by the more intense color in the corresponding part of the plot).

4 Discussion

The exhaustive set of benchmarking experiments allows us to assess the impact of pre-training MLM architectures on historical data for diachronic tasks. As expected, historical pre-training helps to improve performance on diachronically relevant tasks. Accordingly, both TuringBERT and MacBERTh generally outperformed the models pre-trained on contemporary data only—with the exception of both WSD tasks, where MultiBERT (in word-in-context) and BERT (in the parameter-free WSD) outperformed TuringBERT. The historically pre-trained MacBERTh outperformed all competing models across tasks on partitions of the test data stemming from earlier periods.

Furthermore, based on the fact that MacBERTh displays an advantage across tasks and conditions, we can conclude that enlarging the historical span and coverage of pre-training data is advantageous.[15] Importantly, the period to which only MacBERTh had access during pre-training coincides roughly with the beginning of Late Modern English (around the 1700s) and the consolidation of the modern standard. Therefore, if variety in the pre-training data sources results in more powerful feature extractors, the impact for diachronic downstream tasks of the pre-training data stemming from before the 1700s is even so larger.

Interestingly, in many tasks, the advantage shown by MacBERTh was not restricted to the earlier periods. For example, in part-of-speech tagging, the increase in accuracy appeared to be correlated with the total amount of data available for training and testing. This can be interpreted as

[15] Note that this result was obtained even when the total token counts of the pre-training dataset of MacBERTh was smaller with respect to the other models.

a sign of better sample efficiency that pre-training on varied datasets confers the model.

Finally, a question that arises from the present experiments concerns those aspects of the experimental setting that may be responsible for the observed disadvantage of `TuringBERT` on diachronic tasks—even on time spans to which `TuringBERT` had access during pre-training. The fact that the tokenizer is restricted to the specific domain of contemporary English may force the model to aggregate over odd subword tokenizations in order to extract word-level feature representations, putting it in a weakened position. Adapting a model originally pre-trained on contemporary English may also import too strong an inductive bias when the model is later fine-tuned on historical English. In any case, pre-training from scratch on historical data may be a more robust strategy than adapting a pre-trained model.

5 Conclusion & Future Work

Our experiments have shown the potential of historical pre-training for diachronically-relevant tasks. Historical pre-training, however, did not benefit the processing of historical texts from all different time spans to the same extent. A more balanced pre-training dataset could help alleviate these issues. Still, since collecting new data for certain time spans and genres is hindered by the scarcity of such material, researchers are left with the only option of up-sampling the available resources—c.f. Bamman and Burns (2020). The benefits of up-sampling for ranges of the diachrony that are lesser sourced should thus be explored.

Moreover, we have gained insight on the relative merit of different approaches to historical pre-training (pre-training from scratch vs. adapting a pre-existing model). This insight suggests a further experiment in which the "BERT-based Uncased" architecture is fine-tuned on the same dataset as `MacBERTh`, and the resulting model is put to test alongisde `MacBERTh` in order to see whether the claim holds true. However, considering the elevated cost of experimenting with MLM architectures, future research may want to refrain from costly practices like ablation studies and, instead, look at statistical modeling in order to find out the effect of particular design choices—e.g. can excessive sub-word tokenization be responsible for the drop in performance?

Finally, some of the benchmark tasks we implemented were designed ad-hoc to test the capabilities of MLMs at handling historical text. Future work should look into the deployment and evaluation of MLMs in real-word Humanities and DH scenarios in order to scale up the automated retrieval of otherwise difficult to access pieces of information. Besides fine-tuning on appropriate downstream tasks, current NLP research points towards "prompt engineering" (see Liu et al. (2021) for a recent survey) as a promising approach.

References

David Bamman and Patrick J Burns. 2020. Latin BERT: A contextual language model for classical philology.

David Bamman and Gregory Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 1–10.

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. Association for Computational Linguistics.

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

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. 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 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José Moreno, Nicolas Sidère, and Antoine Doucet. 2020. 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.

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.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 288–310. Springer International Publishing.

Susan Fitzmaurice, Justyna A Robinson, Marc Alexander, Iona C Hine, Seth Mehl, and Fraser Dallachy. 2017. Linguistic dna: Investigating conceptual change in early modern english discourse. *Studia Neophilologica*, 89(sup1):21–38.

Lauren Fonteyn. 2020. 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.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. 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. Association for Computational Linguistics.

Edouard Grave. 2017. Language Identification · fastText.

Xiaochuang Han and Jacob Eisenstein. 2019. 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. Association for Computational Linguistics.

Yoonseok Heo, Sangwoo Kang, and Jungyun Seo. 2020. Hybrid sense classification method for large-scale word sense disambiguation. *IEEE Access*, 8:27247–27256.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021a. Neural Language Models for Nineteenth-Century English. *Journal of Open Humanities Data*, 7:22.

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

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.

Leonard Konle and Fotis Jannidis. 2020. 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.

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

British Library Labs. 2014. Digitised books. c. 1510 - c. 1900. json (ocr derived text).

Kai Labusch, Clemens Neudecker, and David Zellhöfer. 2019. 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.

Adam Langley and Dan S. Bloomberg. 2007. Google Books: making the public domain universally accessible. In *Document Recognition and Retrieval XIV*, volume 6500, pages 148 – 157. International Society for Optics and Photonics, SPIE.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 1493–1503. Association for Computational Linguistics.

Jani Marjanen, Lidia Pivovarova, Elaine Zosa, and Jussi Kurunmaki. 2019. 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.

Carlos Martinez-Ortiz, Tom Kenter, Melvin Wevers, Pim Huijnen, and Joris van Eijnatten. 2019. 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.

Martin Mueller, Philip R Burns, and Craig A Berry. 2016. Collaborative curation and exploration of the eebo-tcp corpus. In Laura Estill, Diane K. Jakacki, and Michael Ullyot, editors, *Early Modern Studies after the Digital Turn*, chapter 7, pages 147–167. Iter and the Arizona Center for Medieval and Renaissance Studies.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. 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. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Nils Reimers and Iryna Gurevych. 2019. 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. Association for Computational Linguistics.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. 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.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

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

Stefan Schweter. 2020. Europeana bert and electra models.

Stefan Schweter and Johannes Baiter. 2019. 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. Association for Computational Linguistics.

Stefan Schweter and Luisa März. 2020. 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.

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

Pia Sommerauer and Antske Fokkens. 2019. 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.

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. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. 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.

A Appendix

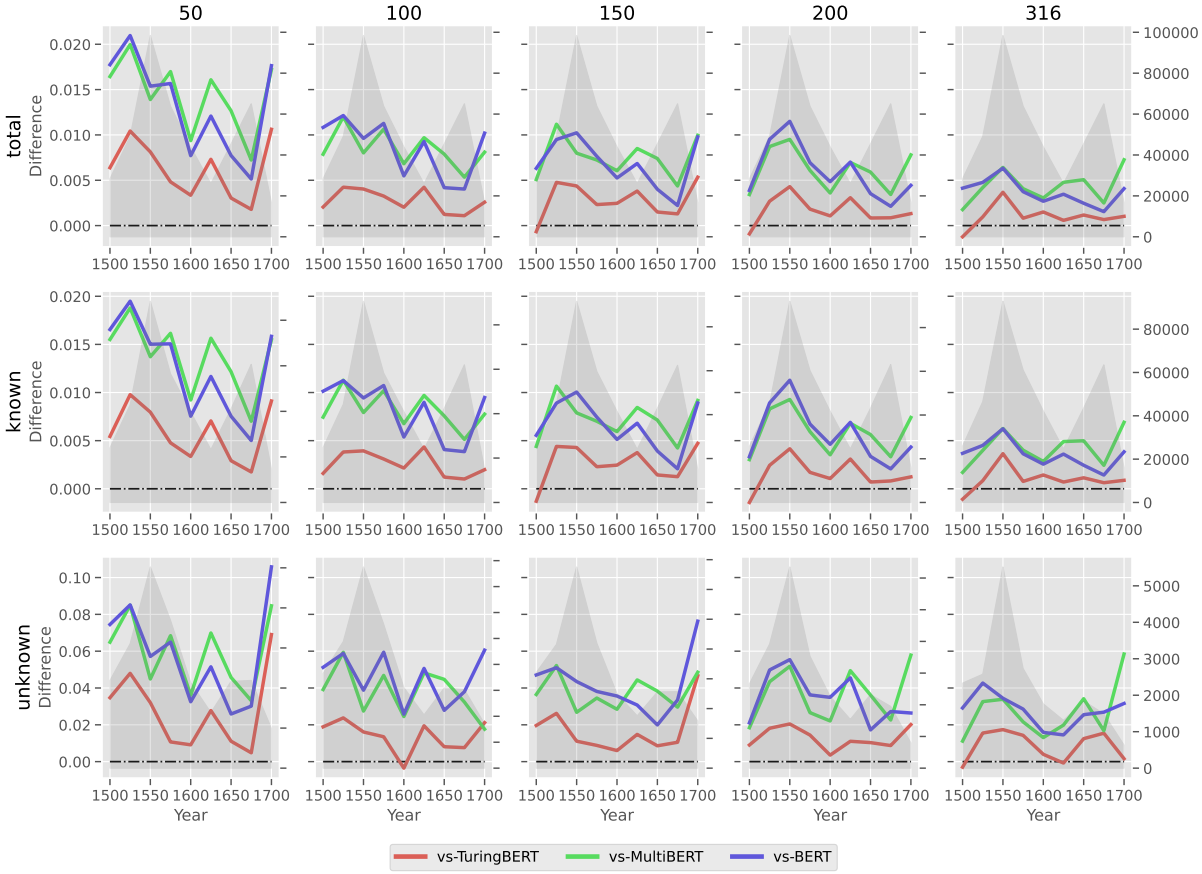

Figure 10: Difference in part-of-speech tagging accuracy of known, unknown and all tokens of `MacBERTh` with respect to the alternative models across training regimes.

	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 3: An example negative pair for lemma "love" showcasing the modification in order to fine-tune the model.

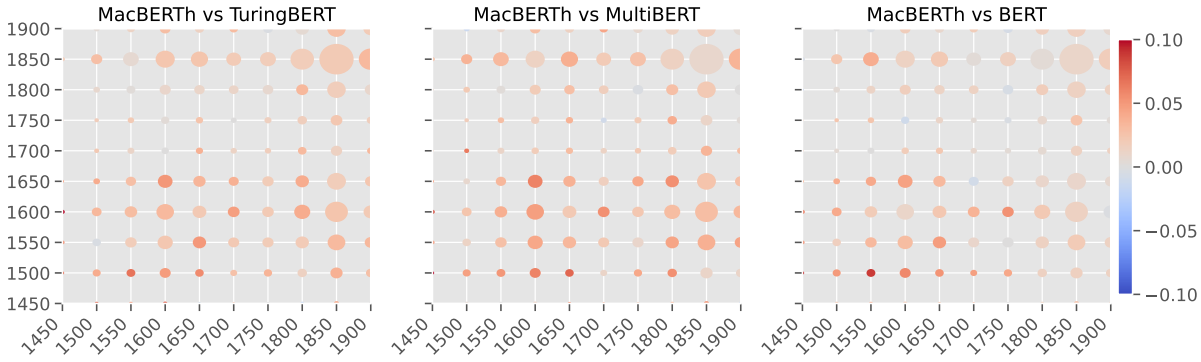

Figure 11: Comparison factoring in the periods of left (y-axis) and right (x-axis) input quotations. Each circle encodes the number (proportional to the radius) as well as the relative difference in accuracy with respect to `MacBERTh` (with red and blue respectively indicating whether `MacBERTh` out- or underperforms the compared models).