Unsupervised Domain Adaptation of Contextualized Embeddings: A Case Study in Early Modern English

Xiaochuang Han\(^1\) and Jacob Eisenstein\(^1,2\)
\(^1\)Georgia Institute of Technology, \(^2\)Facebook AI Research
xc@gatech.edu, me@jacob-eisenstein.com

Abstract
Contextualized word embeddings such as ELMo and BERT provide a foundation for strong performance across a range of natural language processing tasks, in part by pretraining on a large and topically-diverse corpus. However, the applicability of this approach is unknown when the target domain varies substantially from the text used during pretraining. Specifically, we are interested in the scenario in which labeled data is available in only a canonical source domain such as newstext, and the target domain is distinct from both the labeled corpus and the pretraining data. To address this scenario, we propose domain-adaptive fine-tuning, in which the contextualized embeddings are adapted by masked language modeling on the target domain. We test this approach on the challenging domain of Early Modern English, which differs substantially from existing pretraining corpora. Domain-adaptive fine-tuning yields an improvement of 4% in part-of-speech tagging accuracy over a BERT baseline, substantially improving on prior work on this task.\(^1\)

1 Introduction
Contextualized word embeddings have rapidly become a ubiquitous component of natural language processing (Dai and Le, 2015; Devlin et al., 2018; Howard and Ruder, 2018; Radford et al., 2018; Peters et al., 2018). Pretrained contextualized word embeddings can be used as feature for downstream tasks; alternatively, the contextual word embedding module can be incorporated into an end-to-end system, allowing the embeddings to be “fine-tuned” from task-specific labeled data. In either case, a primary benefit of contextualized word embeddings is that they seed the learner with distributional information from large unlabeled datasets.

\(^1\)Trained models will be made available upon publication.

The text used to build pretrained contextualized word embedding models is drawn from a relatively narrow set of domains:

- Wikipedia in BERT (Devlin et al., 2018) and ULMFiT (Howard and Ruder, 2018);
- NewsText (Chelba et al., 2013) in ELMo (Peters et al., 2018);
- BooksCorpus (Zhu et al., 2015) in BERT (Devlin et al., 2018) and GPT (Radford et al., 2018).

All three corpora consist exclusively of text written in the late 20th or 21st centuries; furthermore, Wikipedia and news text are subject to restrictive stylistic constraints (Bryant et al., 2005). It is natural to ask whether the patterns learned from such corpora are transferable to texts from other periods or other stylistic traditions, such as historical documents, technical research papers, and social media.

This paper offers a first empirical investigation of the applicability of domain adaptation to pretrained contextualized word embeddings. As a case study, we focus on the analysis of historical texts. Historical texts are of increasing interest in the computational social sciences and digital humanities, offering insights on patterns of language change (Hilpert and Gries, 2016), social norms (Garg et al., 2018), and the history of ideas and culture (Michel et al., 2011). In particular, researchers have used part-of-speech tagging to identify syntactic changes (Degaetano-Ortlieb, 2018) and dependency parsing to quantify gender-based patterns of adjectival modification and possession in classic literary texts (Vuillemot et al., 2009; Muralidharan and Hearst, 2013). But despite the appeal of using NLP in historical linguistics and literary analysis, there is relatively little research on the impact of diachronic transfer on
the accuracy of taggers and parsers. Indeed, the evidence that does exist suggests that performance degrades significantly, especially if steps are not taken to adapt. For example, Yang and Eisenstein (2015) compare the accuracy of tagging 18th century and 16th century Portuguese text (using a model trained on 19th century test), and find that the error rate doubles.

In this paper, we evaluate the impact of diachronic transfer on a contemporary pretrained part-of-speech tagging system. Our first result is positive: a BERT-based part-of-speech tagger outperforms the state-of-the-art unsupervised domain adaptation method (Yang and Eisenstein, 2016), without taking any explicit steps to adapt to the target domain of Early Modern English. Next, we propose a simple unsupervised domain-adaptive fine-tuning step, using a masked language modeling objective over unlabeled text in the target domain. This yields significant further improvements, with especially strong results on out-of-vocabulary words. Interestingly, domain-adaptive fine-tuning does not decrease performance in the original source domain, yielding a tagger that performs well in both settings.

2 Tagging Historical Texts

Before describing our modeling approach, we highlight some of the unique aspects of tagging historical texts, focusing on the target dataset of the Penn-Helsinki Corpus of Early Modern English (Kroch et al., 2004).

2.1 Early Modern English

Early Modern English (EME) refers to the dominant language spoken in England during the period spanning the 15th-17th centuries, which includes the time of Shakespeare. The English of this era is more comprehensible to contemporary readers than the Middle English that immediately preceded it, but it differs from contemporary English in a number of respects.

From the perspective of natural language processing, a particularly salient aspect of EME is the variability of spelling and other orthographic conventions (Baron and Rayson, 2008). Here is an example:

(1) If this marsch waulle (marsh wall) were not kept, and the canales of eche partes of Sowey river kept from abundance of wedes, al the plaine marsch ground at sodaine raynes (sudden rains) wold be overflowen, and the profite of the meade lost.

While these differences are not too difficult for fluent human readers of English, they affect a large number of tokens, resulting in a substantial increase in the out-of-vocabulary rate (Yang and Eisenstein, 2016). Some of the spelling differences are purely typographical, such as the substitution of v for u in words like vnto, and the substitution of y for i in words like hym. These are particularly common sources of errors for baseline models. Another source of out-of-vocabulary words is the addition of a silent e to the end of many words. This generally did not cause errors for wordpiece-based models, perhaps because the final ‘e’ was segmented as a separate token, which does not receive a tag.

Capitalization is also used inconsistently, making it difficult to distinguish proper and common nouns, e.g.,

(2) And that those Writs which shall be awarded and directed for returning of Ju-ryes . . .

(3) . . . shall not then have Twenty pounds or Eight pounds respectively . . .

Aside from orthography, EME is fairly similar to contemporary English, with a few notable exceptions. Several inflections are rare or non-existent today, such as the -th suffix for third person singular conjugation, as in hath (has) and doth (does). Another difference is in the system of second-person pronouns: EME includes the informal second person thou with associated declensions thee, thine, thy, and the plural second-person pronoun ye. These pronouns are significant sources of errors for baseline models: for example, a BERT-based tagger makes 203 errors on 272 occurrences of the pronoun thee.

2.2 Part-of-Speech Tags in the Penn Parsed Corpora of Historical English

The Penn Parsed Corpora of Historical English (PPCHE) includes part-of-speech tags (Kroch et al., 2004). We focus on the subset of the corpus covering Early Modern English, which we refer to as PPCEME. As discussed in §6, prior work has generally treated tagging the PPCEME as a problem of domain adaptation, with a post-processing stage to map deterministically
between the tagsets. Specifically, we train on the Penn Treebank (PTB) corpus of 20th century English (Marcus et al., 1993), and then evaluate on the PPCEME test set, mapping from the PPCHE tags to the PTB tagset. We use the mapping created by Moon and Baldridge (2007).

One challenge is that the tagsets are difficult to align, particularly with respect to verbs: unlike the PTB, the PPCHE has distinct tags for the modal verbs have, do, and be (and their inflections); unlike the PPCEME, the PTB has distinct tags for third-person singular present indicative (VBZ) and other present indicative verbs (VBP). Moon and Baldridge map only to VBP, and Yang and Eisenstein report an error when VBZ is predicted, even though the corresponding PPCHE tag would be identical in both cases. We avoid this issue by focusing most of our evaluations on a coarse-grained version of the PTB tagset, described in §4.3.

3 Adaptation with Contextualized Embeddings

The problem of processing historical English can be treated as one of unsupervised domain adaptation. Specifically, we assumed that labeled data is available only in contemporary modern English. We now explain how contextualized word embeddings can be applied in this setting.

Contextualized word embeddings provide a generic framework for semi-supervised learning across a range of natural language processing tasks, including sequence labeling tasks like part-of-speech tagging and named entity recognition (Peters et al., 2018; Devlin et al., 2018). Given a sequence of tokens \( w_1, w_2, \ldots, w_T \), these methods return a sequence of vector embeddings \( x_1, x_2, \ldots, x_T \). The embeddings are contextualized, in the sense that they reflect not only each token but also the context in which each token appears. The embedding function is trained either from a language modeling task (Peters et al., 2018) or a related task of recovering masked tokens (Devlin et al., 2018); these methods can be seen as performing semi-supervised learning, because they benefit from large amounts of unlabeled data.

3.1 Task-specific fine-tuning

Recent work has shown that contextualized embeddings are powerful features for a wide range of downstream tasks. Of particular relevance for our work, Devlin et al. show that a state-of-the-art named entity recognition system can be constructed by simply feeding the contextualized embeddings into a linear classification layer. The log probability can then be computed by the log-softmax,

\[
\log p(y_t | w_{1:T}) = \beta_{y_t} \cdot x_t - \log \sum_{y \in Y} \exp (\beta_y \cdot x_t),
\]

where the contextualized embedding \( x_t \) captures information from the entire sequence \( w_{1:T} = (w_1, w_2, \ldots, w_T) \), and \( \beta_y \) is a vector of weights for each tag \( y \).

To fine-tune the contextualized word embeddings, the model is trained by minimizing the negative log-likelihood of the labeled data. This involves backpropagating from the tagging loss into the network that computes contextualized word embeddings. We refer to this procedure as task-fine-tuning in the remainder of the paper.

To borrow from the terminology of domain adaptation (Daumé III and Marcu, 2006), a direct transfer of contextualized word embeddings to the problem of tagging historical text works as follows:

1. Fine-tune BERT for the part-of-speech tagging task, using the Penn Treebank (PTB) corpus of 20th century English;

2. Apply BERT and the learned tagger to the test set of the Penn Parsed Corpus of Early Modern English (PPCEME).

We evaluate this approach in §5

3.2 Unsupervised domain-adaptive fine-tuning

When the target domain differs substantially from the pretraining corpus, the resulting contextual word embeddings may be ineffective for the tagging task. This risk is particularly serious in the setting of unsupervised domain adaptation, in which the labeled data differs substantially from the target text. In this case, task-specific fine-tuning may help adapt the contextualized embeddings towards the labeling task, but not to the domain. To address this issue, we propose the AdaptaBERT model for unsupervised domain adaptation, which adds an additional fine-tuning...
objective: masked language modeling in the target domain. Specifically, we apply the following approach:

1. Randomly mask 15% of tokens in the \textsc{PPCEME} training set, and fine-tune BERT to predict these tokens;
2. Fine-tune the model for the part-of-speech tagging task, using the PTB labels;
3. Apply BERT and the learned tagger to the \textsc{PPCEME} test set.

Step 1 of the method is referred to as \textit{domain-adaptive fine-tuning}, and is based on the same training objective as the original BERT model (Devlin et al., 2018). This step makes it possible to take a pretrained BERT model and adapt it for a target domain, without task labels. Adaptation to the labeling task is then performed in step 2. Attempts to interleave these two steps did not yield significant improvements in performance.

4 Evaluation Setting

We evaluate on the task of part-of-speech tagging in the Penn Parsed Corpus of Early Modern English (\textsc{PPCEME}). There is no canonical training/test split for this data, so we follow Moon and Baldridge in randomly selecting 25% of the documents for the test set. Details of this split are described in Appendix A and in an online code repository.\(^2\)

4.1 Systems

We evaluate the following systems:

\textbf{Frozen BERT} This baseline applies the pretrained “BERT-base” contextualized embeddings, and then learns a tagger from the top-level embeddings. The embeddings are from the pretrained case-sensitive BERT model, and are not adjusted during training.

\textbf{Task-tuned BERT} This baseline starts with pretrained BERT contextualized embeddings, and then fine-tunes them for the part-of-speech tagging task, using the PTB labeled data. This directly follows the methodology for named entity recognition proposed by Devlin et al. (2018) in the original BERT paper.

\textbf{AdaptaBERT} In this approach, we fine-tune the BERT contextualized embedding on the tagging task (with PTB data) and also on a masked language modeling objective on the unlabeled target domain data, as described in § 3. Target domain adaptation is performed on the \textsc{PPCEME} training set.

\textbf{Fine-tuned BERT} This is a \textit{supervised} method, in which we fine-tune the BERT contextualized embeddings on the \textsc{PPCEME} training set. Performance of this method should be viewed as an upper bound, because large-scale labeled data is not available in many domains of interest.

All BERT systems use the pretrained models from Google and the PyTorch implementation from huggingface.\(^3\) Fine-tuning was performed using three NVIDIA Titan XP GPUs. Domain-adaptive fine-tuning took 14.5 hours, and task-tuning took an additional 30 minutes.

4.2 Previous results

We compare the above systems against prior published results from three feature-based taggers:

\textbf{SVM} A support-vector machine baseline tagger, using the surface features described by Yang and Eisenstein (2015).

\textbf{Moon and Baldridge (2007)} This is a logistic regression tagger, using the surface features described by Curran and Clark (2003).

\textbf{FEMA} This is a feature-based unsupervised domain adaptation method for structured prediction (Yang and Eisenstein, 2015), which has the best reported performance on tagging the \textsc{PPCEME}. Unlike AdaptaBERT, the reported results for this system are based on feature induction from the entire \textsc{PPCEME}, including the (unlabeled) test set.

4.3 Tagset mappings

Because we focus on unsupervised domain adaptation, it is not possible to produce tags in the historical English (PPCHE) tagset, which is not encountered at training time. Following

\(^3\)Models retrieved from \url{https://github.com/google-research/bert} on March 14, 2019; implementation retrieved from \url{https://github.com/huggingface/pytorch-pr} also on March 14, 2019.
Moon and Baldridge (2007), we evaluate on a coarsened version of the PTB tagset, using the first letter of each tag (e.g., VBD → v). For comparison with Yang and Eisenstein (2016), we also report results on the full PTB tagset. In these evaluations, the ground truth for the test set is produced by applying the mapping of Moon and Baldridge to the PPCEME tags.

5 Results

As shown in Table 1, fine-tuning to the task and domain each yield significant improvements in performance over the Frozen BERT baseline (line 1). Task-tuning improves accuracy by 7.3% on the coarse-grained tagset (line 2), and domain-adaptive fine-tuning yields a further 3.6% in accuracy (line 3). AdaptaBERT’s performance gains are almost entirely due to a 12.3% improvement on out-of-vocabulary terms, as discussed below.

The rightmost column of the table shows performance on the Penn Treebank test set. Note that domain-adaptive fine-tuning has no impact on the performance on the original tagging task. This shows that adapting the pretrained BERT embeddings to the target domain does not make them less useful for tagging in the source domain, as long as task-tuning is performed after domain-adaptive tuning. In contrast, supervised fine-tuning in the target domain causes performance on the PTB to decrease significantly, as shown in line 4 of the table.

As a secondary evaluation, we measure performance on the full PTB tagset in Table 2, thereby enabling a head-to-head comparison with FEMA (Yang and Eisenstein, 2016). AdaptaBERT outperforms task-tuned BERT by 2.8%, again due to improvements on OOV terms. Task-tuned BERT is on par with the best previous unsupervised domain adaptation result (FEMA), showing the power of contextualized word embeddings, even across disparate domains. Note also that FEMA’s representation was trained on the entire PPCEME, including the test set, while the AdaptaBERT model uses the test set only for evaluation.

5.1 Out-of-vocabulary terms

We define out-of-vocabulary terms as those are not present in the PTB training set. AdaptaBERT’s gains come almost entirely from these terms, with an improvement in OOV accuracy of 24.2% over the frozen BERT baseline and 12.3% over task-tuned BERT. Similarly, on the full PTB tagset, AdaptaBERT attains an improvement in OOV accuracy of 8.6% over FEMA, which was the previous state-of-the-art. These results support our hypothesis that domain-adaptive fine-tuning yields far better contextualized representations for these words. Note that of the out-of-vocabulary words in the PPCEME test set, 52.7% of the types and 82.2% of the tokens appear in the PPCEME training set. This enables domain-adaptive fine-tuning to produce better representations for these terms, making it possible to tag them correctly.

5.2 Errors on in-vocabulary terms

The final two lines of Table 1 indicate that there remains a significant gap between AdaptaBERT and the performance of taggers trained with in-domain data: fine-tuning BERT on the PPCEME training set reduces the error rate to 1.2%. This improvement is largely attributable to in-vocabulary terms: while fine-tuned BERT does better than AdaptaBERT on both IV and OOV terms, the IV terms are far more frequent. Similarly, although AdaptaBERT’s error rate is higher for OOV terms, the largest sources of error are in-vocabulary: the most frequent errors are on tags for to (5337), all (2056), and that (1754); the OOV term with the largest number of errors is vnto (436). These in-vocabulary errors can be explained by inconsistencies in annotation across the two domains:

- In the PPCEME, to may be tagged as either infinitival (to, e.g., I am going to study) or as a preposition (p, e.g., I am going to Italy). However, in the PTB, to is tagged exclusively as TO, which is a special tag reserved for this word.4 Unsupervised domain adaptation generally fails predict the preposition tag for to when it appears in the PPCEME.

- In the PPCEME, all is often tagged as a quantifier (q), which is mapped to adjective (JJ) in the PTB. However, in the PTB, these cases are tagged as determiners (DT), and as a result, the domain adaptation systems always tag all as a determiner.

- In the PTB, the word that is sometimes tagged as a wh-determiner (WDT), in cases such as symptoms that showed up decades

\footnote{In this discussion, \texttt{\textit{sans-serif}} is used for PPCEME tags, and \texttt{\textit{SMALL\ CAPS}} is used for the PTB tags.}
Table 1: Tagging accuracy on PPCEME, using the coarse-grained tagset. The unsupervised systems never see labeled data in the target domain of Early Modern English. The results from Moon and Baldridge (2007) are reprinted from the paper, and only overall accuracy is available. † in line 4, “in-vocab” and “out-of-vocab” refer to the PPCEME training set vocabulary; for lines 1-3, this refers the PTB training set.

Table 2: Tagging accuracy on PPCEME, using the full PTB tagset to compare with Yang and Eisenstein (2016).

later. In the PPCEME, all such cases are tagged as complementizers (c), and this tag is then mapped to the preposition IN. AdaptaBERT often incorrectly tags that as WDT, when IN would be correct.

These examples point to the inherent limitations of unsupervised domain adaptation when there are inconsistencies in the annotation protocol.

6 Related work

BioBERT is an application of BERT to the biomedical domain, which was achieved by pre-training on more than 10 billion tokens of biomedical abstracts and full-text articles from PubMed (Lee et al., 2019). After this pre-training phrase, BioBERT was fine-tuned using task-specific annotations, also in the biomedical domain. A similar approach is employed in SciBERT (Beltagy et al., 2019). In contrast, we do not assume labeled data in the target domain. We therefore adapt pre-trained BERT representation to both the task (part-of-speech tagging) and the domain (Early Modern English).

Universal Language Model Fine-tuning (ULMFiT) also involves fine-tuning on a language modeling task on the target text (Howard and Ruder, 2018). The goal of ULMFiT’s fine-tuning is semi-supervised learning: Howard and Ruder show that accurate text classification can be achieved with fewer labeled examples, but do not consider the issue of domain shift. ULMFiT involves a training regime in which the layers of the embedding model are gradually “unfrozen” during task-specific fine-tuning, to avoid catastrophic forgetting. We do not employ this approach, nor did we experiment with ULMFiT’s elaborate set of learning rate schedules.

Part-of-speech tagging for Early Modern English was explored by Moon and Baldridge (2007), who focused on projecting annotations from labeled out-of-domain data to unlabeled in-domain data. The general problem of adapting part-of-speech tagging over time was considered by Yang and Eisenstein (2015). Their approach projected source (contemporary) and target (historical) training instances into a shared space, by examining the co-occurrence of hand-crafted features. This was shown to significantly reduce the transfer loss in Portuguese, and later in English (Yang and Eisenstein, 2016). However, the approach relies on hand-crafted features, and does not benefit from contemporary neural pre-training architectures, which leverage large-scale unla-
beled data to improve performance across a range of tasks, while making it possible to avoid feature engineering. We find that contemporary neural pre-training yields significant improvements over this previous state-of-the-art.

Another approach to historical text analysis is spelling normalization (e.g., Baron and Rayson, 2008), which has been shown to offer improvements on tagging accuracy (Robertson and Goldwater, 2018). Yang and Eisenstein (2016) found that domain adaptation and normalization were complementary: the performance improvements offered by normalization were orthogonal to those offered by FEMA. In this paper, we have shown that domain-adaptive fine-tuning (and wordpiece segmentation) significantly improves the OOV tagging accuracy from FEMA, so future research must explore whether normalization is still necessary for state-of-the-art tagging of historical texts.

7 Conclusion

This paper demonstrates the applicability of contextualized word embeddings to a difficult unsupervised domain adaptation task, across several centuries of evolution in the English language. We find that BERT works relatively well out-of-the-box, yielding equivalent performance to the best prior unsupervised domain adaptation approaches. Domain-adaptive fine-tuning on unlabeled target domain data yields significant further improvements, especially on OOV terms.

Because many of the differences between contemporary and Early Modern English are orthographic, it is natural to ask whether further gains could be obtained by modeling orthographic changes directly. We plan to investigate this possibility in future work. We are also interested to more thoroughly explore how to combine domain-adaptive and task-specific fine-tuning within the framework of continual learning (Yogatama et al., 2019), with the goal of balancing between these apparently conflicting objectives.

References

Alistair Baron and Paul Rayson. 2008. VarD2: A tool for dealing with spelling variation in historical corpora. In Postgraduate conference in corpus linguistics.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. SciBERT: Pretrained contextualized embeddings for scientific text. arXiv preprint arXiv:1903.10676.

Susan L Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopaedia. In Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work, pages 1–10. ACM.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005.

James R Curran and Stephen Clark. 2003. Investigating GIS and smoothing for maximum entropy taggers. In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), pages 91–98.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In Neural Information Processing Systems (NIPS), pages 3079–3087.

Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research, 26:101–126.

Stefania Degaetano-Ortlieb. 2018. Stylistic variation over 200 years of court proceedings according to gender and social class. In Proceedings of the Second Workshop on Stylistic Variation, pages 1–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16):E3635–E3644.

Martin Hilpert and Stefan Th Gries. 2016. Quantitative approaches to diachronic corpus linguistics. The Cambridge handbook of English historical linguistics, pages 36–53.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In Proceedings of the Association for Computational Linguistics (ACL), pages 328–339.

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME-

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyoo Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: pre-trained biomedical language representation model for biomedical text mining. arXiv preprint arXiv:1901.08746.
Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Taesun Moon and Jason Baldridge. 2007. Part-of-speech tagging for middle english through alignment and projection of parallel diachronic texts. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 390–399.

Aditi Muralidharan and Marti A Hearst. 2013. Supporting exploratory text analysis in literature study. *Literary and linguistic computing*, 28(2):283–295.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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

Alexander Robertson and Sharon Goldwater. 2018. Evaluating historical text normalization systems: How well do they generalize? In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 720–725.

Romain Vuillemot, Tanya Clement, Catherine Plaisant, and Amit Kumar. 2009. What’s being said near “Martha”? Exploring name entities in literary text collections. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 107–114. IEEE.

Yi Yang and Jacob Eisenstein. 2015. Unsupervised multi-domain adaptation with feature embeddings. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical English. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Dani Yogatama, Cyprien de Masson d’Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

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 International Conference on Computer Vision (ICCV)*, pages 19–27.

A Training/test split

No canonical training/test split exists for the PPCEME. We randomly select 25% of the documents for the test set, shown in Table 3 and Table 4. The remaining 75% of documents are used for domain-adaptive fine-tuning in the AdaptaBERT results, and for training in the supervised fine-tuning results.

|             | # documents | # tokens  |
|-------------|-------------|-----------|
| Train       | 333         | 1473103   |
| Test        | 115         | 488054    |

Table 3: Statistics of the training and test set from PPCEME used in our experiments.

---

No canonical training/test split exists for the PPCEME. We randomly select 25% of the documents for the test set, shown in Table 3 and Table 4. The remaining 75% of documents are used for domain-adaptive fine-tuning in the AdaptaBERT results, and for training in the supervised fine-tuning results.
| Document 1 | Document 2 | Document 3 |
|------------|------------|------------|
| alhatton-e3-h.pos | armin-e2-p1.pos | aungier-e3-h.pos |
| aungier-e3-p2.pos | authold-e2-h.pos | bacon-e2-h.pos |
| bacon-e2-p1.pos | blundev-e2-h.pos | boethco-e1-p1.pos |
| boethco-e1-p2.pos | boethpr-e3-p2.pos | burnetroc-e3-p1.pos |
| burnetroc-e3-h.pos | burnetroc-e3-p1.pos | chaplain-e1-p2.pos |
| clowesobs-e2-p2.pos | cromwell-e1-p1.pos | cromwell-e1-p2.pos |
| delapole-e1-p1.pos | deloney-e2-p1.pos | drummond-e3-p1.pos |
| edmonds-e2-h.pos | edmonds-e2-p1.pos | edward-e1-h.pos |
| edward-e1-p1.pos | eliz-1590-e2-p2.pos | elyot-e1-p1.pos |
| eoxinden-1650-e3-p1.pos | fabyan-e1-p1.pos | fhatton-e3-h.pos |
| farquhar-e3-h.pos | fisher-e1-h.pos | forman-diary-e2-p2.pos |
| fiennes-e3-p2.pos | gascoigne-1510-e1-p1.pos | gawdy-e2-h.pos |
| fryer-e3-p1.pos | gawdy-e2-p2.pos | gifford-e2-p1.pos |
| gawdy-e2-p1.pos | harley-e2-h.pos | harley-e2-p1.pos |
| grey-e1-p1.pos | harman-e1-h.pos | henry-1520-e1-h.pos |
| harleyedw-e2-p2.pos | hoole-e3-p2.pos | hoxinden-1640-e3-p1.pos |
| hooker-h-e2-h.pos | jetaylor-e3-h.pos | jethaylor-e3-p1.pos |
| interview-e1-p2.pos | jetaylor-e2-p1.pos | joxinden-e2-p2.pos |
| jopinney-e3-p1.pos | knyvett-1630-e2-p2.pos | koxinden-e2-p1.pos |
| jubarring-e2-p1.pos | kpaston-e2-p1.pos | kscrope-1530-e1-h.pos |
| kpaston-e2-h.pos | leland-e1-p2.pos | lisle-e3-p1.pos |
| leland-e1-h.pos | marches-e1-p1.pos | markham-e2-p2.pos |
| madox-e2-h.pos | masham-e2-p2.pos | memo-e3-p2.pos |
| masham-e2-p1.pos | mroper-e1-p1.pos | mroper-e1-p2.pos |
| miltone=3-h.pos | mroper-e1-p2.pos | penny-e3-p2.pos |
| nhadd-1700-e3-h.pos | perrott-e2-p1.pos | pettit-e2-h.pos |
| pepys-e3-p1.pos | proposals-e3-p2.pos | rceccil-e2-p1.pos |
| pettit-e2-p1.pos | record-e1-p1.pos | record-e1-p2.pos |
| record-e1-h.pos | rhaddsr-1670-e3-p2.pos | rhaddsr-1700-e3-h.pos |
| rhaddsr-1710-e3-p2.pos | roper-e1-h.pos | roxinden-1620-e2-h.pos |
| rplumpt-e1-p1.pos | rplumpt2-e1-p2.pos | shakesp-e2-h.pos |
| shakesp-e2-p1.pos | somers-e3-h.pos | stat-1540-e1-p1.pos |
| stat-1570-e2-p1.pos | stat-1580-e2-p2.pos | stat-1600-e2-h.pos |
| stat-1620-e2-p2.pos | stat-1670-e3-p2.pos | steven-e1-h.pos |
| tillots-a-e3-h.pos | tillots-b-e3-p1.pos | turner-e1-h.pos |
| tyndold-e1-p1.pos | udall-e1-p2.pos | underhill-e1-p2.pos |
| vicary-e1-h.pos | walton-e3-p1.pos | wplumpt-1500-e1-h.pos |
| zouch-e3-p2.pos | | |

Table 4: Test set documents from the PPCEME used in our experiments.