Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling

Xiaochuang Han and **Jacob Eisenstein**
Georgia Institute of Technology*
hxc@cmu.edu, me@jacob-eisenstein.com

Abstract

Contextualized word embeddings such as ELMo and BERT provide a foundation for strong performance across a wide range of natural language processing tasks by pretraining on large corpora of unlabeled text. However, the applicability of this approach is unknown when the target domain varies substantially from the pretraining corpus. We are specifically 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 and pretraining texts. To address this scenario, we propose *domain-adaptive fine-tuning*, in which the contextualized embeddings are adapted by masked language modeling on text from the target domain. We test this approach on sequence labeling in two challenging domains: Early Modern English and Twitter. Both domains differ substantially from existing pretraining corpora, and domain-adaptive fine-tuning yields substantial improvements over strong BERT baselines, with particularly impressive results on out-of-vocabulary words. We conclude that domain-adaptive fine-tuning offers a simple and effective approach for the unsupervised adaptation of sequence labeling to difficult new domains.[1]

1 Introduction

Contextualized word embeddings are becoming a ubiquitous component of natural language processing (Dai and Le, 2015; Devlin et al., 2019; 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 contextualized 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.

However, the texts used to build pretrained contextualized word embedding models are drawn from a narrow set of domains:

- **Wikipedia** in BERT (Devlin et al., 2019) 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., 2019) and GPT (Radford et al., 2018).

All three corpora consist exclusively of text written since the late 20th century; furthermore, Wikipedia and newstext are subject to restrictive stylistic constraints (Bryant et al., 2005).[2] It is therefore crucial to determine whether these pretrained models 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 sequence labeling in historical texts. Historical texts are of increasing interest in the computational social sciences and digital humanities, offering insights on patterns of

*XH is now at Carnegie Mellon University and JE is now at Google Research. Some of the work was performed while JE was visiting Facebook AI Research.

[1]Trained models for Early Modern English and Twitter are available at https://github.com/xhan77/AdaptaBERT

[2]While it might be desirable to completely retrain contextualized word embedding models in the target domain (e.g., Lee et al., 2019), this requires data and computational resources that are often unavailable.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4238–4248,
Hong Kong, China, November 3–7, 2019.

language change (Hilpert and Gries, 2016), social norms (Garg et al., 2018), and the history of ideas and culture (Michel et al., 2011). Syntactic analysis can play an important role: 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 how performance is impacted by diachronic transfer. Indeed, the evidence that does exist suggests that accuracy 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 text), and find that the error rate is twice as high for the older text.

We evaluate the impact of diachronic transfer on a contemporary pretrained part-of-speech tagging system. First, we show that 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 more than 50% error reduction on out-of-vocabulary words. We also present a secondary evaluation on named entity recognition in contemporary social media (Strauss et al., 2016). This evaluation yields similar results: a direct application of BERT outperforms most of the systems from the 2016 WNUT shared task, and domain-adaptive fine-tuning yields further improvements. In both Early Modern English and social media, domain-adaptive fine-tuning does not decrease performance in the original source domain; the adapted tagger performs well in both settings. In contrast, fine-tuning BERT on *labeled* data in the target domain causes catastrophic forgetting (McCloskey and Cohen, 1989), with a significant deterioration in tagging performance in the source domain.

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. While the English of this period is more comprehensible to contemporary readers than the Middle English that immediately preceded it, EME nonetheless differs from the English of today in a number of respects.

Orthography. A particularly salient aspect of EME is the variability of spelling and other orthographic conventions (Baron and Rayson, 2008). For 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 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 (like BERT), perhaps because the final 'e' is 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, as in the following examples:

(2) And that those **Writs** which shall be awarded and directed for returning of **Juryes** ...

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

Morphosyntax. Aside from orthography, EME is fairly similar to contemporary English, with a few notable exceptions. EME includes several inflections that are rare or nonexistent 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 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 216 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) include part-of-speech annotations for texts from several historical periods (Kroch et al., 2004). We focus on the corpus covering Early Modern English, which we refer to as PPCEME. As discussed in § 7, 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, using a mapping between the PPCHE and PTB tagsets defined by Moon and Baldridge (2007).

Unfortunately, there are some fundamental differences in the approaches to verbs taken by each tagset. Unlike the PTB, the PPCHE has distinct tags for the modal verbs *have*, *do*, and *be* (and their infections); unlike the PPCHE, 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 assume that labeled data is available only in the "source domain" of contemporary modern English, and adapt to the "target domain" of historical text by using unlabeled data. 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., 2019). Given a sequence of tokens $w_1, w_2, \ldots, w_T$, these methods return a sequence of vector embeddings $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. 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., 2019); 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

Contextualized embeddings are powerful features for a wide range of downstream tasks. Of particular relevance for our work, Devlin et al. (2019) 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,

$$\log p(y_t \mid \mathbf{w}_{1:T}) = \beta_{y_t} \cdot \mathbf{x}_t - \log \sum_{y \in \mathcal{Y}} \exp\left(\beta_y \cdot \mathbf{x}_t\right), \tag{1}$$

where the contextualized embedding $\mathbf{x}_t$ captures information from the entire sequence $\mathbf{w}_{1:T} = (w_1, w_2, \ldots, w_T)$, and β_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 conditional 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-tuning*.

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 Domain-adaptive fine-tuning

When the target domain differs substantially from the pretraining corpus, the contextualized word embeddings may be ineffective for the tagging task. This risk is particularly serious in unsupervised domain adaptation, because the labeled data may also differ substantially from the target text. In this case, task-specific fine-tuning may help adapt the contextualized embeddings to 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 a simple two-step approach:

1. **Domain tuning**. In the first step, we fine-tune the contextualized word embeddings by backpropagating from the BERT objective, which is to maximize the log-probability of randomly masked tokens.

   We apply this training procedure to a dataset that includes all available target domain data, and an equal amount of unlabeled data in the source domain.[3] We create ten random maskings of each instance; in each masking, 15% of the tokens are randomly masked out, following the original BERT training procedure. We then perform three training iterations over this masked data.

2. **Task tuning**. In the second step, we fine-tune the contextualized word embeddings and learn the prediction model by backpropagating from the labeling objective on the source domain labeled data (Equation 1). This step fine-tunes the contextualized embeddings for the desired labeling task.

Attempts to interleave these two steps did not yield significant improvements in performance. While Peters et al. (2019) report good results on named entity recognition without task tuning, we found this step to be essential in our transfer applications.

[3] If the source domain dataset is smaller than the target domain data, then all of the unlabeled source domain data is included.

	Domain tuning	Task tuning
Prediction	masked tokens	tags
Data	source + target	source

Table 1: Overview of domain tuning and task tuning

4 Evaluation Setting

We evaluate on the task of part-of-speech tagging in the Penn Parsed Corpus of Early Modern English (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 supplement.

4.1 Systems

We evaluate the following systems:

Frozen BERT. This baseline applies the pretrained "BERT-base" contextualized embeddings, and then learns a tagger from the top-level embeddings, supervised by the PTB labeled data. The embeddings are from the pretrained case-sensitive BERT model, and are not adjusted during training. Peters et al. (2019) refer to this as a "feature extraction" application of pretrained embeddings. [4]

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. (2019) in the original BERT paper.

AdaptaBERT. Here we fine-tune the BERT contextualized embedding first on unlabeled data as described in § 3, and then on source domain labeled data. The target domain data is the unlabeled PPCEME training set.

Fine-tuned BERT. In *supervised* learning, we fine-tune the BERT contextualized embeddings on the labeled PPCEME training set.

[4] Note that this baseline learns only a linear final layer over the pretrained embeddings. We do not adopt weighted combination of internal contextual layers or any non-linear final layers in Peters et al. (2019).

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.[5] Fine-tuning was performed using one NVIDIA GeForce RTX 2080 TI GPU. Domain-adaptive fine-tuning took 12 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:

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

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 PPCEME. Unlike AdaptaBERT, the reported results for this system are based on feature induction from the entire 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 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 is produced by applying the mapping of Moon and Baldridge to the PPCEME tags.

5 Results

Fine-tuning to the task and domain each yield significant improvements in performance over the Frozen BERT baseline (Table 2, line 1). Task-tuning improves accuracy by 7.6% on the coarse-grained tagset (line 2), and domain-adaptive fine-tuning yields a further 4.5% in accuracy (line 3). AdaptaBERT's performance gains are almost entirely due to the improvement on out-of-vocabulary terms, as discussed below.

The rightmost column of the table shows performance on the Penn Treebank test set. Interestingly, 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. This can be viewed as a form of *catastrophic forgetting* (McCloskey and Cohen, 1989).

As a secondary evaluation, we measure performance on the full PTB tagset in Table 3, thereby enabling direct comparison with prior work (Yang and Eisenstein, 2016). AdaptaBERT outperforms task-tuned BERT by 3.9%, 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 unlabeled 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 that are not present in the PTB training set. Of the out-of-PTB-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. Indeed, AdaptaBERT's gains come almost entirely from these terms, with an improvement in OOV accuracy of 25.8% over the frozen BERT baseline and 15.7% over task-tuned BERT. Similarly, on the full PTB tagset, AdaptaBERT attains an improvement in OOV accuracy of 11.3% over FEMA, which was the previous state-of-the-art.

5.2 Errors on in-vocabulary terms

The final two lines of Table 2 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 train-

[5] Models retrieved from https://github.com/google-research/bert on March 14, 2019; implementation retrieved from https://github.com/huggingface/pytorch-pretrained-bert also on March 14, 2019.

	Early Modern English			PTB
System	Accuracy	In-vocab	Out-of-vocab	Accuracy
Unsupervised domain adaptation				
1. Frozen BERT	77.7	83.7	61.0	91.4
2. Task-tuned BERT	85.3	90.4	71.1	98.2
3. AdaptaBERT (this work)	**89.8**	**90.8**	**86.8**	**98.2**
Supervised in-domain training				
4. Fine-tuned BERT	98.8	99.0†	93.2†	92.4

Table 2: Tagging accuracy on PPCEME, using the coarse-grained tagset. The unsupervised systems never see labeled data in the target domain of Early Modern English. † 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.

System	Accuracy	In-vocab	Out-of-vocab
1. SVM	74.2	81.7	49.9
2. FEMA (Yang and Eisenstein, 2016)	77.9	82.3	63.2
3. Task-tuned BERT	78.4	84.4	58.4
4. AdaptaBERT (this work)	**82.3**	**84.7**	**74.5**

Table 3: Tagging accuracy on PPCEME, using the full PTB tagset to compare with Yang and Eisenstein (2016).

ing set reduces the error rate to 1.2%. This improvement is largely attributable to *in-vocabulary* terms: although fine-tuned BERT does better than AdaptaBERT on both IV and OOV terms, the IV terms are far more frequent. The most frequent errors for AdaptaBERT are on tags for *to* (5337), *all* (2054), and *that* (1792). The OOV term with the largest number of errors is *al* (306), which is a shortening of *all*.

These errors on in-vocabulary terms 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.[6] 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 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 Social Media Microblogs

As an additional evaluation, we consider social media microblogs, of which Twitter is the best known example in English. Twitter poses some of the same challenges as historical text: orthographic variation leads to a substantial mismatch in vocabulary between the target domain and source training documents such as Wikipedia (Baldwin et al., 2013; Eisenstein, 2013). We hypothesize that domain-adaptive fine-tuning can help to produce better contextualized word embeddings for microblog text.

Our evaluation is focused on the problem of identifying named entity spans in Tweets, which was the shared task of the 2016 Workshop on Noisy User Text (WNUT; Strauss et al., 2016). In

[6] In this discussion, sans-serif is used for PPCEME tags, and SMALL CAPS is used for the PTB tags.

the shared task, systems were given access to labeled data in the target domain; in contrast, we are interested to measure whether it is possible to perform this task without access to such data. As training data, we use the canonical CoNLL 2003 shared task dataset, in which named entity spans were annotated on a corpus of newstext (Tjong Kim Sang and De Meulder, 2003). Because the entity types are different in the WNUT and CoNLL corpora, we focus on the *segmentation task* of identifying named entity spans. Participants in the 2016 WNUT shared task competed on this metric, enabling a direct comparison with the performance of these supervised systems.

Results are shown in Table 4. A baseline system using task-tuned BERT (trained on the CoNLL labeled data) achieves an F1 of 57.7 (line 1). This outperforms six of the ten submissions to the WNUT shared task, even though these systems are trained on in-domain data. AdaptaBERT yields marginal improvements when domain-adaptive fine-tuning is performed on the WNUT training set (line 2); expanding the target domain data with an additional million unlabeled tweets yields a 2.3% improvement over the BERT baseline (line 3). Performance improves considerably when the domain-adaptive fine-tuning is performed on the combined WNUT training and test sets (line 4).

Test set adaptation is controversial: it is not realistic for deployed systems that are expected to perform well on unseen instances without retraining, but it may be applicable in scenarios in which we desire good performance on a predefined set of documents. The WNUT test set was constructed by searching for tweets in a narrow time window on two specific topics: shootings and cybersecurity events. It is therefore unsurprising that test set adaptation yields significant improvements, since it can yield useful representations of the names of the relevant entities, which might not appear in a random sample of tweets. This is a plausible approach for researchers who are interested in finding the key entities participating in such events in an pre-selected corpus of text.

When BERT is fine-tuned on labeled data in the target domain, the performance of the resulting tagger improves to 64.3% (line 5). This would have achieved second place in the 2016 WNUT shared task. The state-of-the-art system makes use of character-level information that is not available to our models (Limsopatham and Collier, 2016). As with the evaluation on Early Modern English, we find that domain-adaptive fine-tuning does not impair performance on the source domain (CoNLL), but supervised in-domain training increases the source domain error rate dramatically.

7 Related Work

Adaptation in neural sequence labeling. Most prior work on adapting neural networks for NLP has focused on supervised domain adaptation, in which a labeled data is available in the target domain (Mou et al., 2016). RNN-based models for sequence labeling can be adapted across domains by manipulating the input or output layers individually (e.g., Yang et al., 2016) or simultaneously (Lin and Lu, 2018). Unlike these approaches, we tackle unsupervised domain adaptation, which assumes only unlabeled instances in the target domain. In this setting, prior work has focused on *domain-adversarial* objectives, which construct an auxiliary loss based on the capability of an adversary to learn to distinguish the domains based on a shared encoding of the input (Ganin et al., 2016; Purushotham et al., 2017). However, adversarial methods require balancing between at least two and as many as six different objectives (Kim et al., 2017), which can lead to instability (Arjovsky et al., 2017) unless the objectives are carefully balanced (Alam et al., 2018).

In addition to supervised and unsupervised domain adaptation, there are "distantly supervised" methods that construct noisy target domain instances, e.g., by using a bilingual dictionary (Fang and Cohn, 2017). Normalization dictionaries exist for Early Modern English (Baron and Rayson, 2008) and social media (Han et al., 2012), but we leave their application to distant supervision for future work.

Finally, language modeling objectives have previously been used for domain adaptation of text classifiers (Ziser and Reichart, 2018), but this prior work has focused on representation learning from scratch, rather than adaptation of a pretrained contextualized embedding model. Our work shows that models that are pretrained on large-scale data yield very strong performance even when applied out-of-domain, making them a natural starting point for further adaptation.

Semi-supervised learning. Universal Language Model Fine-tuning (ULMFiT) also involves fine-tuning on a language modeling task on the target

	WNUT				CoNLL
System	domain adaptation data	Precision	Recall	**F1**	F1
Unsupervised domain adaptation					
1. Task-tuned BERT	n/a	50.9	66.6	57.7	**97.8**
2. AdaptaBERT	WNUT training	52.8	66.7	58.9	97.6
3. AdaptaBERT	+ 1M tweets	53.6	68.3	60.0	97.8
4. AdaptaBERT	WNUT train+test	57.7	68.9	**62.8**	97.8
Supervised in-domain training					
5. Fine-tuned BERT	n/a	66.3	62.3	64.3	80.9
6. Limsopatham and Collier (2016)	n/a	73.5	59.7	**65.9**	

Table 4: Named entity segmentation performance on the WNUT test set and CoNLL test set A. Limsopatham and Collier (2016) had the winning system at the 2016 WNUT shared task. Their results are reprinted from their paper, which did not report performance on the CoNLL dataset.

text (Howard and Ruder, 2018), but the goal is semi-supervised learning: prior work shows that accurate text classification can be achieved with fewer labeled examples, but does 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-tuning, to avoid catastrophic forgetting. We do not employ this approach, nor did we experiment with ULMFiT's elaborate set of learning rate schedules. Finally, contemporaneous unpublished work has applied the BERT objective to target domain data, but like ULMFiT, this work focuses on semi-supervised classification rather than cross-domain sequence labeling (Xie et al., 2019).

Tagging historical texts. Moon and Baldridge (2007) approach part-of-speech tagging of Early Modern English by 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, this approach relies on hand-crafted features, and does not benefit from contemporary neural pretraining architectures. We show that pretrained contextualized embeddings yield significant improvements over feature-based methods.

Normalization. Another approach to historical texts and social media is spelling normalization (e.g., Baron and Rayson, 2008; Han et al., 2012), which has been shown to offer improvements in tagging historical texts (Robertson and Goldwater, 2018). In Early Modern English, Yang and Eisenstein (2016) found that domain adaptation and normalization are complementary. 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.

Domain-specific pretraining. Given sufficient unlabeled data in the target domain, the simplest approach may be to retrain a BERT-like model from scratch. BioBERT is an application of BERT to the biomedical domain, which was achieved by pretraining on more than 10 billion tokens of biomedical abstracts and full-text articles from PubMed (Lee et al., 2019); SciBERT is a similar approach for scientific texts (Beltagy et al., 2019). Data on this scale is not available in Early Modern English, but retraining might be applicable to social media text, assuming the user has access to both large-scale data and sufficient computing power. The mismatch between pretrained contextualized embeddings and technical NER corpora was explored in recent work by Dai et al. (2019).

8 Conclusion

This paper demonstrates the applicability of contextualized word embeddings to two difficult un-

supervised domain adaptation tasks. On the task of adaptation to historical text, BERT works relatively well out-of-the-box, yielding equivalent performance to the best prior unsupervised domain adaptation approach. Domain-adaptive fine tuning on unlabeled target domain data yields significant further improvements, especially on OOV terms. On the task of adaptation to contemporary social media, a straightforward application of BERT yields competitive results, and domain-adaptive fine tuning again offers improvements.

A potentially interesting side note is that while supervised fine-tuning in the target domain results in *catastrophic forgetting* of the source domain, unsupervised target domain tuning does not. This suggests the intriguing possibility of training a single contextualized embedding model that works well across a wide range of domains, genres, and writing styles. However, further investigation is necessary to determine whether this finding is dependent on specific details of the source and target domains in our experiments, or whether it is a general difference between unsupervised domain tuning and supervised fine-tuning. 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.

Acknowledgments. Thanks to the anonymous reviewers and to Ross Girshick, Omer Levy, Michael Lewis, Yuval Pinter, Luke Zettlemoyer, and the Georgia Tech Computational Linguistics Lab for helpful discussions of this work. The research was supported by the National Science Foundation under award RI-1452443.

References

Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain adaptation with adversarial training and graph embeddings. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1077–1087.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 214–223.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364.

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 encyclopedia. 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*.

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

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using similarity measures to select pretraining data for NER. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 359–369.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(59):1–35.

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.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 421–432.

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.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Adversarial adaptation of synthetic or stale data. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1297–1307.

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Nut Limsopatham and Nigel Collier. 2016. Bidirectional LSTM for named entity recognition in twitter messages. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 145–152, Osaka, Japan. The COLING 2016 Organizing Committee.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

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.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

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.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in NLP applications? In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 479–489.

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)*.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Sanjay Purushotham, Wilka Carvalho, Tanachat Nilanon, and Yan Liu. 2017. Variational recurrent adversarial deep domain adaptation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pages 142–147.

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

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation. *arXiv preprint arXiv:1904.12848*.

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)*.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2016. Transfer learning for sequence tagging with hierarchical recurrent networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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.

Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1241–1251.