Part-of-Speech Tagging for Historical English

Yi Yang and **Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA 30308
{yiyang+jacobe}@gatech.edu

Abstract

As more historical texts are digitized, there is interest in applying natural language processing tools to these archives. However, the performance of these tools is often unsatisfactory, due to language change and genre differences. Spelling normalization heuristics are the dominant solution for dealing with historical texts, but this approach fails to account for changes in usage and vocabulary. In this empirical paper, we assess the capability of domain adaptation techniques to cope with historical texts, focusing on the classic benchmark task of part-of-speech tagging. We evaluate several domain adaptation methods on the task of tagging Early Modern English and Modern British English texts in the Penn Corpora of Historical English. We demonstrate that the Feature Embedding method for unsupervised domain adaptation outperforms word embeddings and Brown clusters, showing the importance of embedding the entire feature space, rather than just individual words. Feature Embeddings also give better performance than spelling normalization, but the combination of the two methods is better still, yielding a 5% raw improvement in tagging accuracy on Early Modern English texts.

1 Introduction

There is growing interest in applying natural language processing (NLP) techniques to historical texts (Piotrowski, 2012), with applications in information retrieval (Dougherty, 2010; Jurish, 2011), linguistics (Baron et al., 2009; Rayson et al., 2007), and the digital humanities (Hendrickx et al., 2011; Muralidharan and Hearst, 2013; Pettersson and Nivre, 2011). However, these texts differ from contemporary training corpora in a number of linguistic respects, including the lexicon (Giusti et al., 2007), morphology (Borin and Forsberg, 2008), and syntax (Eumeridou et al., 2004). This imposes significant challenges for modern NLP tools: for example, the accuracy of the CLAWS part-of-speech Tagger (Garside and Smith, 1997) drops from 97% on the British National Corpus to 82% on Early Modern English texts (Rayson et al., 2007). There are two main approaches that could improve the accuracy of NLP systems on historical texts: normalization and domain adaptation.

Original: and drewe vnto hym all ryottours & wylde dysposed persones
Normalization: and drew unto him all ryottours & wild disposed persons

Figure 1: An example sentence from Early Modern English and its VARD normalization.

Normalization Spelling normalization (also called canonicalization) involves mapping historical spellings to their canonical forms in modern languages, thus bridging the gap between contemporary training corpora and target historical texts. Figure 1 shows one historical sentence and its normalization by VARD (Baron and Rayson, 2008). Rayson et al. (2007) report an increase of about 3% accuracy on adaptation of POS tagging from Modern English texts to Early Modern English texts if the target texts were automatically normalized by the VARD system. However, normalization is not always a well-defined problem (Eisenstein,

Proceedings of NAACL-HLT 2016, pages 1318–1328,
San Diego, California, June 12-17, 2016.

2013), and it does not address the full range of linguistic changes over time, such as unknown words, morphological differences, and changes in the meanings of words (Kulkarni et al., 2015). In the example above, the word 'ryottours' is not successfully normalized to 'rioters'; the syntax is comprehensible to contemporary English speakers, but usages such as 'wild disposed' and 'drew unto' are sufficiently unusual as to pose problems for NLP systems trained on contemporary texts.

Domain adaptation A more generic machine learning approach is to apply unsupervised domain adaptation techniques, which transform the representations of the training and target texts to be more similar, typically using feature co-occurrence statistics (Blitzer et al., 2006; Ben-David et al., 2010). It is natural to think of historical texts as a distinct domain from contemporary training corpora, and Yang and Eisenstein (2014, 2015) show that the accuracy of historical Portuguese POS tagging can be significantly improved by domain adaption. However, we are unaware of prior work that empirically evaluates the efficacy of this approach on Early Modern English texts. Furthermore, historical texts are often associated with multiple metadata attributes (e.g., author, genre, and epoch), each of which may influence the text's linguistic properties. *Multi-domain adaptation* (Mansour et al., 2009) and *multi-attribute domain adaptation* (Joshi et al., 2013; Yang and Eisenstein, 2015) can potentially exploit these metadata attributes to obtain further improvements.

This paper presents the first comprehensive empirical comparison of effectiveness of these approaches for part-of-speech tagging on historical texts. We focus on the two historical treebanks of the Penn Corpora of Historical English — the Penn Parsed Corpus of Modern British English (Kroch et al., 2010, PPCMBE) and the Penn-Helsinki Parsed Corpus of Early Modern English (Kroch et al., 2004, PPCEME). These datasets enable a range of analyses, which isolate the key issues in dealing with historical corpora:

- In one set of analyses, we focus on the PPCMBE and the PPCEME corpora, training on more recent texts and testing on earlier texts. This isolates the impact of language change on tagging performance.
- In another set of analyses, we train on the Penn Treebank (Marcus et al., 1993, PTB), and test on the historical corpora, using the tag mappings from Moon and Baldridge (2007). We apply the well-known Stanford CoreNLP tagger to this task (Manning et al., 2014), thus replicating the most typical situation for users of existing language technology.
- We show that FEMA, a domain adaptation algorithm that is specifically designed for sequence labeling problems (Yang and Eisenstein, 2015), achieves an increase of nearly 4% in tagging accuracy when adapting from the PTB to the PPCEME.
- We compare the impact of normalization with domain adaptation, and demonstrate that they are largely complementary.
- Error analysis shows that the improvements obtained by domain adaptation are largely due to better handling of out-of-vocabulary (OOV) tokens. Many of the most frequent errors on in-vocabulary (IV) tokens are caused by mismatches in the tagsets or annotation guidelines, and may be difficult to address without labeled data in the target domain.

2 Data

The Penn Corpora of Historical English consist of the Penn-Helsinki Parsed Corpus of Middle English, second edition (Kroch et al., 2010, PPCME2), the Penn-Helsinki Parsed Corpus of Early Modern English (Kroch et al., 2004, PPCEME), and the Penn Parsed Corpus of Modern British English (Kroch and Taylor, 2000, PPCMBE). The corpora are annotated with part-of-speech tags and syntactic parsing trees in an annotation style similar to that of the Penn Treebank. In this work, we focus on POS tagging the PPCMBE and the PPCEME.[1]

[1] Middle English is outside the scope of this paper, because it is sufficiently unintelligible to modern English speakers that texts such as Canterbury Tales are published in translation. In tagging Middle English texts, Moon and Baldridge (2007) apply bitext projection techniques from multilingual learning, rather than domain adaptation.

Period	# Sentence	# Token
1840-1914	17,770	322,255
1770-1839	23,462	427,424
1700-1769	16,083	343,024
Total	57,315	1,092,703

Table 1: Statistics of the Penn Parsed Corpus of Modern British English (PPCMBE), by time period.

Period	# Sentence	# Token
1640-1710	29,181	614,315
1570-1639	39,799	706,587
1500-1569	31,416	640,255
Total	100,396	1,961,157

Table 2: Statistics of the Penn Parsed Corpus of Early Modern English (PPCEME), by time period.

The Penn Parsed Corpus of Modern British English The PPCMBE is a syntactically annotated corpus of text, containing roughly one million word tokens from documents written in the period 1700-1914. It is divided into three 70-year time periods according to the composition date of the works. Table 1 shows the statistics of the corpus by time period.[2] In contrast to the PTB, the PPCMBE contains text from a variety of genres, such as Bible, Drama, Fiction, and Letters.

The Penn-Helsinki Parsed Corpus of Early Modern English The PPCEME is a collection of text samples from the Helsinki Corpus (Rissanen et al., 1993), as well as two supplements mainly consisting of text material by the same authors and from the same editions as the material in the Helsinki Corpus. The corpus contains nearly two million words from texts in the period from 1500 until 1710, and it is divided into three 70-year time periods similar to the PPCMBE corpus. The statistics of the corpus by time period is summarized in Table 2. The PPCEME consists of text from the same eighteen genres as the PPCMBE.

Penn Treebank Release 3 The Penn Treebank (Marcus et al., 1993) is the de facto standard syntactically annotated corpus for English, which is used to train software such as Stanford CoreNLP (Manning et al., 2014). When using this dataset for supervised training, we follow Toutanova et al. (2003) and use WSJ sections 0-18 for training, and sections 19-21 for tuning. When applying unsupervised domain adaptation, we use all WSJ sections, together with texts from the PPCMBE and the PPCEME.

Tagsets The Penn Corpora of Historical English (PCHE) use a tagset that differs from the Penn Treebank, mainly in the direction of greater specificity. Auxiliary verbs 'do', 'have', and 'be' all have their own tags, as do words like 'one' and 'else', due to their changing syntactic function over time. Overall, there are 83 tags in the PPCEME, and 81 in the PPCMBE, as compared with 45 in the PTB. Furthermore, the tags in the PCHE tagset are allowed to join constituent morphemes in compounds, yielding complex tags such as `PRO+N` (e.g., 'himself') and `ADJ+NS` (e.g., 'gentlemen').

To measure the tagging accuracy of PTB-trained taggers on the historical texts, we follow Moon and Baldridge (2007), who define a set of deterministic mappings from the PCHE tags to the PTB tagset. For simplicity, we first convert each complex tag to the simple form by only considering the first simple tag component (e.g., `PRO+N` to `PRO` and `ADJ+NS` to `ADJ`). This has little effect on the tagging performance, as the complex tags cover only slightly more than 1% of the tokens in the PCHE treebanks. Among the 83 tags, 74 mappings to the corresponding PTB tags are obtained from Moon and Baldridge (2007). We did our best to convert the other tags according to the tag description. The complete list of mappings is published in Appendix A.

3 Unsupervised Domain Adaptation

In typical usage scenarios, the user wants to tag some historical text but has no labeled data in the target domain (e.g., Muralidharan and Hearst, 2013). This best fits the paradigm of unsupervised domain adaptation, when labeled data from the source domain (e.g., the PTB) is combined with unlabeled data from the target domain. Representational differences between source and target domains can be a major source of errors in domain adaptation (Ben-

[2] All the statistics in this section include punctuation, but exclude extra-linguistic material such as page numbers or token ID numbers.

David et al., 2010), and so several representation learning approaches have been proposed.

The most straightforward approach is to replace lexical features with **word representations**, such as Brown clusters (Brown et al., 1992; Lin et al., 2012) or word embeddings (Turian et al., 2010), such as word2vec (Mikolov et al., 2013). Lexical features can then be replaced or augmented with the resulting word representations. This can assist in domain adaptation by linking out-of-vocabulary words to invocabulary words with similar distributional properties.

Word representations are suitable for adapting lexical features, but a more general solution is to adapt the entire feature representation. One such method is **Structural Correspondence Learning** (Blitzer et al., 2006, **SCL**). In SCL, we create artificial binary classification problems for thousands of cross-domain "pivot" features, and then use the weights from the resulting classifiers to project the instances into a new dense representation. We also consider a recently-published approach called **Feature Embedding** (**FEMA**), which achieves the state-of-the-art results on several POS tagging adaptation tasks (Yang and Eisenstein, 2015). The intuition of FEMA is similar to SCL and other prior work: it relies on co-occurrence statistics to link features across domains. Specifically, FEMA exploits the tendency of many NLP tasks to divide features into templates, and induces feature embeddings by using the features in each template to predict the active features in all other templates — just as the skipgram model learns word embeddings to predict neighboring words. The resulting embeddings can be substituted for the "one-hot" representation of each feature template, resulting in a dense, low-dimensional representation of each instance.

A further advantage of FEMA is that it can perform multi-attribute domain adaptation, enabling it to exploit the many metadata attributes (e.g., year, genre, and author) that are often associated with historical texts. This is done by accounting for the specific impact of each domain attribute on the feature predictors, and then building a domain-neutral representation from the common substructure that is shared across all domain attributes. In the experiments that follow, we use genre and epoch as domain attributes.

4 Experiments

We evaluate these unsupervised domain adaptation approaches on part-of-speech tagging for historical English (the PPCMBE and the PPCEME), in two settings: (1) temporal adaptation within each individual corpus, where we train POS taggers on the most modern data in the corpus and test on increasingly distant datasets; (2) adaptation of English POS tagging from modern news text to historical texts. The first setting focuses on temporal differences, and eliminates other factors that may impair tagging performance, such as different annotation schemes and text genres. The second setting is the standard and well-studied evaluation scenario for POS tagging, where we train on the Wall Street Journal (WSJ) text from the PTB and test on historical texts. In addition, we evaluate the effectiveness of the VARD normalization tool (Baron and Rayson, 2008) for improving POS tagging performance on the PPCEME corpus.

4.1 Experimental Settings

The datasets used in the experiments are described in § 2. All the hyperparameters are tuned on development data in the source domain. In the case where there is no specific development dataset (adaptation within the historical corpora), we randomly sample 10% sentences from the training datasets for hyperparameter tuning.

4.1.1 Baseline systems

We include two baseline systems for POS tagging: a classification-based support vector machine (SVM) tagger and a bidirectional maximum entropy Markov model (MEMM) tagger. Specifically, we use the L_2-regularized L_2-loss SVM implementation in the scikit-learn package (Pedregosa et al., 2011) and L_2-regularized bidirectional MEMM implementation provided by Stanford CoreNLP (Toutanova et al., 2003; Manning et al., 2014).

Following Yang and Eisenstein (2015), we apply the feature templates defined by Ratnaparkhi (1996) to extract the basic features for all taggers. There are three broad types of templates: five lexical feature templates, eight affix feature templates, and three orthographic feature templates.

Task	baseline		SCL	Brown	word2vec	FEMA	
	SVM	MEMM (Stanford)				single embedding	attribute embeddings (error reduction)
Modern British English (training from 1840-1914)							
→ 1770-1839	96.30	96.57	96.42	96.45	96.44	96.80	**96.84** (15%)
→ 1700-1769	94.57	94.83	95.07	95.15	94.85	95.65	**95.75** (22%)
AVERAGE	95.43	95.70	95.74	95.80	95.64	96.23	**96.30** (19%)
Early Modern English (training from 1640-1710)							
→ 1570-1639	93.62	93.98	94.23	94.36	94.18	95.01	**95.20** (25%)
→ 1500-1569	87.59	87.47	89.39	89.73	89.30	91.40	**91.63** (33%)
AVERAGE	90.61	90.73	91.81	92.05	91.74	93.20	**93.41** (30%)

Table 3: Accuracy results for temporal adaptation in the PPCMBE and the PPCEME of historical English. Percentage error reduction is shown for the best-performing method, FEMA-attribute.

4.1.2 Domain adaptation systems

We consider the unsupervised domain adaptation methods described in § 3: structural correspondence learning (SCL), Brown clustering, word2vec,[3] and FEMA, which we train in both the single embedding mode (FEMA-single), where metadata attributes are ignored, and in multi-attribute mode (FEMA-attribute), where metadata attributes are used. The domain adaptation models are trained on the union of the (unlabeled) source and target datasets. This ensures that there are no out-of-vocabulary items for the word or feature embeddings.

Following Yang and Eisenstein (2015), we do not learn feature embeddings for the three orthographic feature templates: as each orthographic feature template represents only a binary value, it is unnecessary to replace it with a much longer numerical vector. The learned representations are then concatenated with the basic surface features to form the augmented representations. For computational reasons, the domain adaptation systems are all based on the SVM tagger, as pilot studies showed that Viterbi tagging offers minimal improvements.

4.1.3 Parameter tuning

We choose the SVM regularization parameter by sweeping the range $\{0.1, 0.3, 0.5, 0.8, 1.0\}$. Following Blitzer et al. (2006), we consider pivot features that appear more than 50 times in all the domains for SCL. We empirically fix the number of singular vectors of the projection matrix K to 25, and also employ feature normalization and rescaling, as these settings yield best performance in prior work. The number of Brown clusters is chosen from the range $\{50, 100, 200, 400\}$. For FEMA and word2vec, we choose embedding sizes from the range $\{50, 100, 200, 300\}$ and fix the numbers of negative samples to 15. The window size for training word embeddings is set as 5. Finally, we adopt the same regularization penalty for all the attribute-specific embeddings of FEMA, which is selected from the range $\{0.01, 0.1, 1.0, 10.0\}$. All parameters were tuned on development data in the source domain. We train the Stanford MEMM tagger using the default configuration file.

[3] https://code.google.com/p/word2vec/

4.2 Temporal Adaptation

In the temporal adaptation setting, we work within each corpus, training on the most recent section, and evaluating on the two earlier sections. For PPCMBE, the source domain is the period from 1840 to 1914; for PPCEME, the source domain is the period from 1640 to 1710. All earlier texts are treated as target domains. We transform the tags to the PTB tagset for evaluation, so that results can be compared with the next experiment, in which the PTB is used for supervision.

Settings We randomly sample 10% sentences from the training data as the development data for optimizing hyperparameters, and then retrain the models on the full training data using the best parameters. For FEMA, we consider domain attributes

for 70-year temporal periods and genres, resulting in a total of 21 attributes for each corpus. The numbers of pivot features used in SCL are 4400 and 5048 for the PPCMBE and the PPCEME respectively. The best number of Brown clusters is 200, and the best embedding sizes are 200 and 100 for word2vec and FEMA.

Results As shown in Table 3, accuracies are significantly improved by domain adaptation, especially for the PPCEME. English spelling had become mostly uniform and stable since around 1700 (Baron et al., 2009), which may explain why improvements on the PPCMBE are relatively modest, especially in the 1770-1839 epoch. Among the two baseline systems, MEMM performs slightly better than SVM, showing a small benefit to structured prediction. Among the domain adaptation algorithms, FEMA clearly outperforms SCL, Brown clustering and word2vec, with an averaged increase of about 0.5% and 1.5% accuracies on the PPCMBE and the PPCEME test sets respectively. The metadata attribute information boosts performance by a small but consistent margin, 0.1-0.2% on average.

4.3 Adaptation from the Penn Treebank

Newspaper text is the primary data source for training modern NLP systems. For example, most "off-the-shelf" English POS taggers (e.g., the Stanford Tagger (Toutanova et al., 2003), SVMTool (Giménez and Marquez, 2004), and CRFTagger (Phan, 2006)) are trained on the WSJ portion of the Penn Treebank, which is composed of professionally-written news text from 1989. This motivates this evaluation scenario, in which we train the tagger on the Penn Treebank WSJ data and apply it to historical English texts, using all sentences of the PPCMBE and PPCEME for testing.

Settings The feature representations are trained on the union of the PTB and the PPCEME. The domain attributes for FEMA are set to include the three corpora themselves (PTB, PPCMBE, and PPCEME), and the genre attributes in the historical corpora. Note that all sentences in the Penn Treebank WSJ data belong to the same genre (news). For SCL, we use the same threshold of 50 occurrences for pivot features, and include 8089 features that pass this threshold. PTB WSJ sections 19-21 are used for parameter tuning: we find that the best number of Brown clusters is 200, and the optimum embedding sizes are 200 and 100 for word2vec and FEMA.

Spelling normalization Spelling variants lead to a high percentage of out-of-vocabulary (OOV) tokens in historical texts, which poses problems for POS tagging. We normalize the PPCEME sentences using VARD (Baron and Rayson, 2008), a widely used spelling normalization tool that has been proven to improve performance on POS tagging (Rayson et al., 2007) and syntactic parsing (Schneider et al., 2014). VARD is designed specifically for Early Modern English spelling variation, and additional labeled data and training are required for other forms of spelling variation, which we do not consider here. Following Schneider et al. (2014), we utilize VARD's auto-normalization function with a 50% normalization threshold, achieving a balance between precision and recall. At this threshold, a total of 12% ($236298/1961157$) of the tokens in the PPCEME are normalized.[4]

Results As shown in Table 4, this task is considerably more difficult, with even the best systems achieving accuracies that are nearly 15% worse than in-domain training. Nonetheless, domain adaptation can help: FEMA improves performance by 1.3% on the PPCMBE data, and by 3.8% on the unnormalized PPCEME data. Spelling normalization also helps, improving the baseline systems by more than 2.5%. The combination of spelling normalization and domain adaptation gives an overall improvement in accuracy from 74.2% to 79.1%. The relative error reduction is lower than in the temporal adaptation setting: only 19% at best, versus 30% error reduction in temporal adaptation. This is because there are now at least two sources of error — language change and tagset mismatch — and unsupervised domain adaptation cannot address mismatches in the tag set.

5 Analysis

As expected, the Early Modern English dataset (PPCEME) is considerably more challenging than the Modern British English dataset (PPCMBE): the

[4] We only consider $1:1$ mappings, and ignore 328 normalizations corresponding to $1:n$ mappings.

Target	Normalized	baseline		SCL	Brown	word2vec	FEMA	
		SVM	MEMM (Stanford)				single embedding	attribute embeddings (error reduction)
PPCMBE	No	81.12	81.35	81.66	81.65	81.75	82.34	**82.46** (7%)
PPCEME	No	74.15	74.34	75.89	76.04	75.85	77.77	**77.92** (15%)
PPCEME	Yes	76.73	76.87	77.61	77.65	77.76	78.85	**79.05** (19%*)

Table 4: Accuracy results for adapting from the PTB to the PPCMBE and the PPCEME of historical English. *Error reduction for the normalized PPCEME is computed against the unnormalized SVM accuracy, showing total error reduction.

baseline accuracy is 7% worse on the PPCEME than the PPCMBE. However, the PPCEME is also more amenable to domain adaptation, with FEMA offering considerably larger improvements. One reason is that the PPCEME has many more out-of-vocabulary (OOV) tokens: 23%, versus 9.2% in the PPCMBE. Both domain adaptation and normalization help to address this specific issue, and they yield further improvements when used in combination. This section offers further insights on the sources of errors and possibilities for improvement on the PPCEME data.

Feature set	IV	OOV	All
All features	81.68	48.96	74.15
– word context	79.69	38.62	70.23
– prefix	81.61	46.11	73.43
– suffix	81.36	38.13	71.40
– affix	81.22	34.40	70.44
– orthographic	81.68	48.92	74.14

Table 5: Tagging accuracies of adaptation of our baseline SVM tagger from the PTB to the PPCEME in ablation experiments.

5.1 Feature Ablation

Table 5 presents the results of feature ablation experiments for the non-adapted SVM tagger. Word context features are important for obtaining good accuracies on both IV and OOV tokens. Affix features, particularly suffix features, are crucial for the OOV tokens. The orthographic features are shown to be nearly irrelevant, as long as affix features are present. Overall, the high percentage of OOV tokens can be a major source of errors, as the tagging accuracy on OOV tokens is below 50% in our best baseline system. Note that these results are for a classification-based tagger; while the Viterbi-based MEMM tagger performs only marginally better overall ($\sim 0.2\%$ improvement), it is possible that its error distribution might be different due to the advantages of structured prediction.

5.2 Error Analysis

The accuracy on out-of-vocabulary (OOV) tokens is generally low, and spelling variation is a major source of OOV tokens. For instance, 'ye' and 'thy', the older forms of 'the' and 'your', are often incorrectly tagged as `NN` and `JJ` in the PPCEME. In general, the per-tag accuracies are roughly correlated with the percentages of OOV tokens. Some exceptions including `VB`, `NNP` and `NNS`, where the affix features can be very useful for tagging OOV tokens.

That said, the cross-domain accuracy on in-vocabulary (IV) tokens is also low, at roughly 80% when adapting from the PTB to the PPCEME. A major source of error here is the mismatch in annotation schemes between the two datasets, which is only partially addressed by a deterministic tag mapping. Table 6 presents the SVM accuracy per tag, and the most common error correspondingly. Most of the errors shown in the table are owing to different annotations of the same token in the two corpora.

One major cause of errors is in misalignments of punctuations and their POS tags. For example, in the PPCEME, 16.6% of commas are labeled as `.` (sentence-final punctuation), and 12.3% periods are labeled as `,` (sentence-internal punctuation); these punctuations are less ambiguous in the PTB. The historical corpora lack special tags for colons and ellipses, which are present in the PTB. In contrast to the PTB, there is no distinction between opening quotation mark and closing quotation mark in the PPCEME. Moon and Baldridge (2007) avoid these difficulties by mapping all the punctuation tokens

Tag	% of OOV	Accuracy	Most common error
IN	6.93	82.79	to/TO
NN	48.39	64.74	Lord/NNP
DT	3.45	94.62	that/IN
PRP	13.57	78.80	other/JJ
,	0.41	87.86	./.
JJ	32.20	48.60	all/DT
CC	1.98	91.29	for/IN
RB	26.22	65.74	such/JJ
.	0.56	54.43	,/,
VB	34.69	75.06	have/VBP
NNP	58.91	88.31	god/NN
NNS	59.12	73.88	Lords/NNPS
VBD	25.87	81.93	quoth/NN
VBN	37.75	63.09	said/VBD
PRP$	13.57	85.49	thy/JJ

Table 6: Accuracy (recall) rates per tag with the SVM model, for the 15 most common tags. For each gold category, the most common error word and predicted tag are shown.

to a single tag. We did not follow their setting because it would lead to a significant change of test data. However, it should be noted that these "errors" are not particularly meaningful for linguistic analysis, and could easily be addressed by heuristic post-processing.

The tagging performance is also impaired by the different annotations of many common words. For example, in the PTB, more than 99.9% of token 'to' are labeled as `TO`, but in the PCHE this word can also be labeled as `IN`, distinguishing the infinitive marker from the preposition. The words 'all', 'any' and 'every' are annotated as quantifiers in the PCHE; this tag is mapped to `JJ`, but these specific words are all labeled as `DT` in the PTB. A simple remapping from `Q` to `DT` leads to an increase of 0.78% baseline accuracy; it is possible that other changes to the tag mappings of Moon and Baldridge (2007) might yield further improvements, but a more systematic approach would be outside the bounds of *unsupervised* domain adaptation.

5.3 Improvements from Normalization

As shown above, the tagging accuracy decreases from 81.7% on IV tokens to 49.0% on OOV tokens. Spelling normalization helps to increase the accuracy by transforming OOV tokens to IV tokens. After normalization, the OOV rate for the PPCEME falls from 23.0% to 13.5%, corresponding to a reduction of 41.5% OOV tokens. Normalization is not perfectly accurate, and the tagging performance for IV tokens drops slightly to 81.2% on IV tokens. But due to the dramatic decrease in the number of OOV tokens, normalization improves the overall accuracy by more than 2.5%. We also observe performance drops on tagging OOV tokens after normalization (49.0% to 48.1%), which suggests that the remaining unnormalized OOV tokens are the tough cases for both normalization and POS tagging.

System	IV	OOV	All
SVM	81.68	48.96	74.15
SCL	82.01	55.45	75.89
Brown	81.81	56.76	76.04
word2vec	81.79	56.00	75.85
FEMA-single	82.30	62.63	77.77
FEMA-attribute	82.34	63.16	77.92

Table 7: Tagging accuracies of domain adaptation models from the PTB to the PPCEME.

5.4 Improvements from Domain Adaptation

As presented in Table 7, the tagging accuracies are increased on both IV and OOV tokens with the domain adaptation methods. Compared against the baseline tagger, FEMA-attribute achieves an absolute improvement of 14% in accuracy on OOV tokens. SCL performs slightly better than Brown clustering and word2vec on IV tokens, but worse on OOV tokens. By incorporating metadata attributes, FEMA-attribute performs better than FEMA-single on OOV tokens, though the accuracies on IV tokens are similar. Interestingly, the venerable method of Brown clustering (slightly) outperforms both word2vec and SCL.

We further study the relationship between domain adaptation and spelling normalization by looking into the errors corrected by both approaches. Domain adaptation yields larger improvements than spelling normalization on both IV and OOV tokens, although as noted above, the approaches are somewhat complementary. The results show that among the 60,928 error tokens corrected by VARD, 60% are also corrected by FEMA-attribute, while the remaining 40% would be left uncorrected by the domain

adaptation technique. Conversely, among the errors corrected by FEMA-attribute, 38% are also corrected by VARD, while the remaining 62% would be left uncorrected. The overlap of reduced errors is because both approaches exploit similar sources of information, including affixes and local word contexts.

6 Related Work

Domain adaptation Early work on domain adaptation focuses on supervised setting, in which some amount of labeled instances are available in the target domain (Jiang and Zhai, 2007; Daumé III, 2007; Finkel and Manning, 2009). Unsupervised domain adaptation is more challenging but attractive in many applications, and several representation learning methods have been proposed for addressing this problem. Structural Correspondence Learning (Blitzer et al., 2006, SCL) and marginalized denoising autoencoders (Chen et al., 2012, mDA) seek cross-domain representations that are useful to predict a subset of features in the original instances, called pivot features. Schnabel and Schütze (2014) directly induce distributional representations for POS tagging based on local left and right neighbors of the token. More recent work trains cross-domain representations with neural networks, with additional objectives such as minimizing errors in the source domain and maximizing domain confusion loss (Ganin and Lempitsky, 2015; Tzeng et al., 2015). We show the Feature Embedding model, which is specifically designed for NLP problems with feature templates (Yang and Eisenstein, 2015), achieves strong performance on historical adaptation tasks.

Historical texts Historical texts differ from modern texts in spellings, syntax and semantics, posing significant challenges for standard NLP systems, which are usually trained with modern news text. Numerous resources have been created for overcoming the difficulties, including syntactically annotated corpora (Kroch et al., 2004; Kroch et al., 2010; Galves and Faria, 2010) and spelling normalization tools (Giusti et al., 2007; Baron and Rayson, 2008). Most previous work focuses on normalization, which can significantly increase tagging accuracy on historical English (Rayson et al., 2007) and German (Scheible et al., 2011). Similar improvements have been obtained for syntactic parsing (Schneider et al., 2014). Domain adaptation offers an alternative approach which is more generic — for example, it can be applied to any corpus without requiring the design of a set of normalization rules. As shown above, when normalization is possible, it can be combined with domain adaptation to yield better performance than that obtained by either approach alone.

7 Conclusion

Syntactic analysis is a key first step towards processing historical texts, but it is confounded by changes in spelling and usage over time. We empirically evaluate several unsupervised domain adaptation approaches for POS tagging of historical English texts. We find that domain adaptation methods significantly improve the tagger performance on two historical English treebanks, with relative error reductions of 30% in the temporal adaptation setting. FEMA outperforms other domain adaptation approaches, showing the importance of adapting the entire feature vector, rather than simply using word embeddings. Normalization and domain adaptation combine to yield even better performance, with a total of 5% raw accuracy improvement over a baseline classifier in the most difficult setting. Error analysis reveals that tagset mismatch is the most common source of errors for in-vocabulary words. We hope that our work encourages further research on domain adaptation for historical texts and provides useful baselines in these efforts.

Acknowledgments This research was supported by National Science Foundation award 1349837, and by the National Institutes of Health under award number R01GM112697-01. We thank the reviewers for their helpful feedback.

References

Alistair Baron and Paul Rayson. 2008. Vard2: A tool for dealing with spelling variation in historical corpora. In *Postgraduate conference in corpus linguistics*.

Alistair Baron, Paul Rayson, and Dawn Archer. 2009. Word frequency and key word statistics in corpus linguistics. *Anglistik*, 20(1):41–67.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman

Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 120–128.

Lars Borin and Markus Forsberg. 2008. Something old, something new: A computational morphological description of old swedish. In *LREC 2008 workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 9–16.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Minmin Chen, Z. Xu, Killian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the Association for Computational Linguistics (ACL)*, Prague.

William C Dougherty. 2010. The Google Books Project: will it make libraries obsolete? *The Journal of Academic Librarianship*, 36(1):86–89.

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, Atlanta, GA.

Eugenia Eumeridou, Blaise Nkwenti-Azeh, and John McNaught. 2004. An analysis of verb subcategorization frames in three special language corpora with a view towards automatic term recognition. *Computers and the Humanities*, 38(1):37–60.

Jenny R. Finkel and Christopher Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 602–610, Boulder, CO.

Charlotte Galves and Pablo Faria. 2010. Tycho Brahe Parsed Corpus of Historical Portuguese. `http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html`.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Roger Garside and Nicholas Smith. 1997. A hybrid grammatical tagger: Claws4. *Corpus annotation: Linguistic information from computer text corpora*, pages 102–121.

Jesús Giménez and Lluis Marquez. 2004. SVMTool: A general POS tagger generator based on support vector machines. In *Proceedings of the Language Resources and Evaluation Conference*.

Rafael Giusti, A Candido, Marcelo Muniz, Lívia Cucatto, and Sandra Aluísio. 2007. Automatic detection of spelling variation in historical corpus. In *Proceedings of the Corpus Linguistics Conference (CL)*.

Iris Hendrickx, Michel Généreux, and Rita Marquilhas. 2011. Automatic pragmatic text segmentation of historical letters. In *Language Technology for Cultural Heritage*, pages 135–153. Springer.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the Association for Computational Linguistics (ACL)*, Prague.

Mahesh Joshi, Mark Dredze, William W. Cohen, and Carolyn P. Rosé. 2013. What's in a domain? multi-domain learning for multi-attribute data. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 685–690, Atlanta, GA.

Bryan Jurish. 2011. *Finite-state canonicalization techniques for historical German*. Ph.D. thesis, Universitätsbibliothek.

Anthony Kroch and Ann Taylor. 2000. Penn-Helsinki Parsed Corpus of Middle English, second edition. `http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html`.

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`.

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2010. The Penn Parsed Corpus of Modern British English. `http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html`.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of International Conference on World Wide Web (WWW)*, pages 625–635.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky.

2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2009. Domain adaptation with multiple sources. In *Neural Information Processing Systems (NIPS)*, pages 1041–1048.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems (NIPS)*, pages 3111–3119, Lake Tahoe.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Eva Pettersson and Joakim Nivre. 2011. Automatic verb extraction from historical swedish texts. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 87–95.

Xuan-Hieu Phan. 2006. CRFTagger: CRF English POS Tagger. `http://crftagger.sourceforge.net`.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 133–142.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the bard: Evaluating the accuracy of a modern pos tagger on early modern english corpora. In *Corpus Linguistics Conference*.

Matti Rissanen, Merja Kytö, and Minna Palander-Collin. 1993. *Early English in the computer age: Explorations through the Helsinki Corpus*. Number 11. Walter de Gruyter.

Silke Scheible, Richard J Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS-tagger on early modern German text. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pages 19–23.

Tobias Schnabel and Hinrich Schütze. 2014. Flors: Fast and simple domain adaptation for part-of-speech tagging. *Transactions of the Association of Computational Linguistics*, 2:51–62.

Gerold Schneider, Hans Martin Lehmann, and Peter Schneider. 2014. Parsing early and late modern english corpora. *Literary and Linguistic Computing*.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word Representation: A Simple and General Method for Semi-Supervised Learning. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 384–394, Uppsala, Sweden.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4068–4076.

Yi Yang and Jacob Eisenstein. 2014. Fast easy unsupervised domain adaptation with marginalized structured dropout. In *Proceedings of the Association for Computational Linguistics (ACL)*, Baltimore, MD.

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)*, Denver, CO.