Part-of-speech Tagging for Middle English through Alignment and Projection of Parallel Diachronic Texts

Taesun Moon and Jason Baldridge
Department of Linguistics
University of Texas at Austin
1 University Station B5100
Austin, TX 78712-0198 USA
tsmoon, jbaldrid@mail.utexas.edu

Abstract

We demonstrate an approach for inducing a tagger for historical languages based on existing resources for their modern varieties. Tags from Present Day English source text are projected to Middle English text using alignments on parallel Biblical text. We explore the use of multiple alignment approaches and a bigram tagger to reduce the noise in the projected tags. Finally, we train a maximum entropy tagger on the output of the bigram tagger on the target Biblical text and test it on tagged Middle English text. This leads to tagging accuracy in the low 80's on Biblical test material and in the 60's on other Middle English material. Our results suggest that our bootstrapping methods have considerable potential, and could be used to semi-automate an approach based on incremental manual annotation.

1 Introduction

Annotated corpora of historical texts provide an important resource for studies of syntactic variation and change in diachronic linguistics. For example, the Penn-Helsinki Parsed Corpus of Middle English (PPCME) (Kroch and Taylor, 2000) has been used to show the existence of syntactic dialectal differences between northern and southern Middle English (Kroch et al., 2000) and to examine the syntactic evolution of the English imperative construction (Han, 2000). However, their utility rests on their having coverage of a significant amount of annotated material from which to draw patterns for such studies, and creating resources such as the PPCME require significant time and cost to produce. Corpus linguists interested in diachronic language studies thus need efficient ways to produce such resources.

One approach to get around the annotation bottleneck is to use semi-automation. For example, when producing part-of-speech tags for the Tycho Brahe corpus of Historical Portuguese (Britto et al., 2002), a set of seed sentences was manually tagged, and the Brill tagger (Brill, 1995) was then trained on those and consequently used to tag other sentences. The output was inspected for errors, the tagger was retrained and used again to tag new sentences, for several iterations.

We also seek to reduce the human effort involved in producing part-of-speech tags for historical corpora. However, our approach does so by leveraging existing resources for a language's modern varieties along with parallel diachronic texts to produce accurate taggers. This general technique has worked well for bilingual bootstrapping of language processing resources for one language based on already available resources from the other. The first to explore the idea were Yarowsky and Ngai (2001), who induced a part-of-speech tagger for French and base noun phrase detectors for French and Chinese via transfer from English resources. They built a highly accurate POS tagger by labeling English text with an existing tagger (trained on English resources), aligning that text with parallel French, projecting the automatically assigned English POS tags across these alignments, and then using the automatically labeled French text to train a new French tagger. This tech-

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 390–399, Prague, June 2007.

nique has since been used for other languages and tasks, e.g. morphological analysis (Yarowsky et al., 2001), fine-grained POS tagging for Czech (Drábek and Yarowsky, 2005), and tagging and inducing syntactic dependencies for Polish (Ozdowska, 2006).

This methodology holds great promise for producing tools and annotated corpora for processing diachronically related language pairs, such as Modern English to Middle or Old English. Historical languages suffer from a paucity of machine readable text, inconsistencies in orthography, and grammatical diversity (in the broadest sense possible). This diversity is particularly acute given that diachronic texts of a given language encompass texts and genres spanning across centuries or millenia with a plethora of extra-linguistic influences to complicate the data. Furthermore, even in historically contemporaneous texts, possible dialectal variations further amplify the differences in already idiosyncratic orthographies and syntactic structure.

The present study goes further than Britto et al. (2002) by fully automating the alignment, POS tag induction, and noise elimination process. It is able to utilize the source language to a greater degree than the previously mentioned studies that attempted language neutrality; that is, it directly exploits the genetic similarity between the source and target language. Some amount of surface structural similarity between a diachronic dialect and its derivatives is to be expected, and in the case of Middle English and Modern English, such similarities are not negligible.

The automation process is further aided through the use of two versions of the Bible, which obviates the need for sentence alignment. The modern Bible is tagged using the C&C maximum entropy tagger (Curran and Clark, 2003), and these tags are transferred from source to target through high-confidence alignments aquired from two alignment approaches. A simple bigram tagger is trained from the resulting target texts and then used to relabel the same texts as Middle English training material for the C&C tagger. This tagger utilizes a rich set of features and a wider context, so it can exploit surface similarities between the source and target language. By training it with both the original (Modern English) Penn Treebank Wall Street Journal (WSJ) material and our automatically tagged Middle English Wycliffe material, we achieve an accuracy of 84.8% on predicting coarse tags, improving upon a 63.4% baseline of training C&C on the WSJ sentences alone. Furthermore, we show that the bootstrapped tagger greatly reduces the error rate on out-of-domain, non-Biblical Middle English texts.

2 Data

English provides an ideal test case for our study because of the existence of publically accessible diachronic texts of English and their translations in electronic format and because of the availability of the large, annotated Penn-Helsinki Parsed Corpus of Middle English. The former allows us to create a POS tagger via alignment and projection; the latter allows us to evaluate the tagger on large quantities of human-annotated tags.

2.1 The Bible as a parallel corpus

We take two versions of the Bible as our parallel corpus. For modern English, we utilize the NET Bible[1]. For Middle English (ME), we utilize John Wycliffe's Bible[2]. The first five lines of Genesis in both Bibles are shown in Figure 1.

The Bible offers some advantages beyond its availability. All its translations are numbered, facilitating assessment of accuracy for sentence alignment models. Also, the Bible is quite large for a single text: approximately 950,000 words for Wycliffe's version and 860,000 words for the NET bible. Finally, Wycliffe's Bible was released in the late 14th century, a period when the transition of English from a synthetic to analytical language was finalized. Hence, word order was much closer to Modern English and less flexible than Old English; also, nominal case distinctions were largely neutralized, though some verbal inflections such as distinctions for the first and second person singular in the present tense were still in place (Fennell, 2001). This places Wycliffe's Bible as far back as possible without introducing extreme nominal and verbal inflections in word alignment.

The two Bibles were cleaned and processed for the present task and then examined for levels of correspondence. The two texts were compared for

[1]The New English Translation Bible, which may be downloaded from http://www.bible.org/page.php?page_id=3086.

[2]Available for download at: http://wesley.nnu.edu/biblical_studies/wycliffe.

1 In the beginning God created the heavens and the earth.
2 Now the earth was without shape and empty, and darkness was over the surface of the watery deep, but the Spirit of God was moving over the surface of the water.
3 God said, "Let there be light." And there was light!
4 God saw that the light was good, so God separated the light from the darkness.
5 God called the light day and the darkness night. There was evening, and there was morning, marking the first day.

1 In the bigynnyng God made of nouyt heuene and erthe.
2 Forsothe the erthe was idel and voide, and derknessis weren on the face of depthe; and the Spiryt of the Lord was borun on the watris.
3 And God seide, Liyt be maad, and liyt was maad.
4 And God seiy the liyt, that it was good, and he departide the liyt fro derknessis; and he clepide the liyt,
5 dai, and the derknessis, nyyt. And the euentid and morwetid was maad, o daie.

Figure 1: The first five verses of Genesis the NET Bible (top) and Wycliffe's Bible (below).

whether there were gaps in the chapters and whether one version had more chapters over the other. If discrepancies were found, the non-corresponding chapters were removed. Next, because we assume sentences are already aligned in our approach, discrepancies in verses between the two Bibles were culled. A total of some two hundred lines were removed from both Bibles. This processing resulted in a total of 67 books[3], with 920,000 words for the Wycliffe Bible and 840,000 words for the NET Bible.

2.2 The Penn-Helsinki Parsed Corpus of Middle English

The Penn-Helsinki Parsed Corpus of Middle English is a collection of text samples derived from manuscripts dating 1150–1500 and composed during the same period or earlier. It is based on and expands upon the Diachronic Part of the Helsinki Corpus of English Texts. It contains approximately 1,150,000 words of running text from 55 sources. The texts are provided in three forms: raw, POS tagged, and parsed.

Among the texts included are portions of the Wycliffe Bible. They comprise partial sections of *Genesis* and *Numbers* from the Old Testament and *John I.1–XI.56* from the New Testament. In total, the sections of Wycliffe annotated in PPCME have some 25,000 words in 1,845 sentences. This was used as part of the test material. It is important to note that there are significant spelling differences from the full Wycliffe text that we use for alignment – this is a common issue with early writings that makes building accurate taggers for them more difficult than for the clean and consistent, edited modern texts typically used to evaluate taggers.

2.3 Tagsets

The PPCME uses a part-of-speech tag set that has some differences from that used for the Penn Treebank, on which modern English taggers are generally trained. It has a total of 84 word tags compared to the widely used Penn Treebank tag set's 36 word tags.[4] One of the main reasons for the relative diversity of the PPCME tag set is that it maintains distinctions between the *do*, *have*, and *be* verbs in addition to non-auxiliary verbs. The tag set is further complicated by the fact that composite POS tags are allowed as in *another_D+OTHER*, *midnyght_ADJ+N*, or *armholes_N+NS*.

To measure tagging accuracy, we consider two different tag sets: PTB, and COARSE. A measurement of accuracy is not possible with a direct comparison to the PPMCE tags since our approach la-

[3] 66 books shared by the churches and one book from the Apocrypha. A comparison of the two Bibles revealed that the NET Bible contained the Apocrypha, but only Baruch was shared between the two versions.

[4] In our evaluations, we collapse the many different punctuation tags down to a single tag, $PUNC$.

bels target text in Middle English with tags from the Penn Treebank. Therefore, with PTB, all non-corresponding PPCME tags were conflated if necessary and mapped to the Penn Treebank tag set. Between the two sets, only 8 tags, EX, FW, MD, TO, VB, VBD, VBN, VBP, were found to be fully identical. In cases where tags from the two sets denoted the same category/subcategory, one was simply mapped to the other. When a PPCME tag made finer distinctions than a related Penn tag and could be considered a subcategory of that tag, it was mapped accordingly. For example, the aforementioned auxiliary verb tags in the PPMCE were all mapped to corresponding subcategories of the larger VB tag group, a case in point being the mapping of the perfect participle of *have*_HVN to VBN, a plain verbal participle. For COARSE, the PTB tags were even further reduced to 15 category tags,[5] which is still six more than the core consensus tag set used in Yarkowsky and Ngai (2001). Specifically, COARSE was measured by comparing the first letter of each tag. For example, *NN* and *NNS* are conflated to *N*.

2.4 Penn Treebank Release 3

The POS tagged Wall Street Journal, sections 2 to 21, from the Penn Treebank Release 3 (Marcus et al., 1994) was used to train a Modern English tagger to automatically tag the NET Bible. It was also used to enhance the maximum likelihood estimates of a bigram tagger used to label the target text.

3 Approach

Our approach involves three components: (1) projecting tags from Modern English to Middle English through alignment; (2) training a bigram tagger; and (3) bootstrapping the C&C tagger on Middle English texts tagged by the bigram tagger. This section describes these components in detail.

3.1 Bootstrapping via alignment

Yarowsky and Ngai (2001) were the first to propose the use of parallel texts to bootstrap the creation of taggers. The approach first requires an alignment to be induced between the words of the two texts; tags are then projected from words of the source language to words of the target language. This naturally leads to the introduction of noise in the target language tags. Yarowsky and Ngai deal with this by (a) assuming that each target word can have at most two tags and interpolating the probability of tags given a word between the probabilities of the two most likely tags for that word and (b) interpolating between probabilities for tags projected from 1-to-1 alignments and those from 1-to-n alignments. Each of these interpolated probabilities is parameterized by a single variable; however, Yarowsky and Ngai do not provide details for how the two parameter values were determined/optimized.

Here, we overcome much of the noise by using two alignment approaches, one of which exploits word level similarities (present in genetically derived languages such as Middle English and Present Day English) and builds a bilingual dictionary between them. We also fill in gaps in the alignment by using a bigram tagger that is trained on the noisy tags and then used to relabel the entire target text.

The C&C tagger (Curran and Clark, 2003) was trained on the Wall Street Journal texts in the Penn Treebank and then used to tag the NET Bible (the source text). The POS tags were projected from the source to the Wycliffe Bible based on two alignment approaches, the Dice coefficient and Giza++, as described below.

3.1.1 Dice alignments

A dictionary file is built using the variation of the Dice Coefficient (Dice (1945)) used by Kay and Röscheisen (1993):

$$D(v,w) = \frac{2c}{N_A(v) + N_B(w)} \geq \theta$$

Here, c is the number of cooccurring positions and $N_T(x)$ is the number of occurrences of word x in corpus T. c is calculated only once for redundant occurrences in an aligned sentence pair. For example, it is a given that *the* will generally occur more than once in each aligned sentence. However, even if *the* occurs more than once in each of the sentences in aligned pair s_A and s_B, c is incremented only once. v and w are placed in the word alignment table if they exceed the threshold value θ, which is an empirically determined, heuristic measure.

[5] Namely, adjective, adverb, cardinal number, complementizer/preposition, conjunction, determiner, existential *there*, foreign word, interjection, infinitival *to*, modal, noun, pronoun, verb, and *wh*-words.

The dictionary was structured to establish a surjective relation from the target language to the source language. Therefore, no lexeme in the Wycliffe Bible was matched to more than one lexeme in the NET Bible. The Dice Coefficient was modified so that for a given target word v

$$D_v = \arg\max_w D(v, w)$$

would be mapped to a corresponding word from the source text, such that the Dice Coefficient would be maximized. Dictionary entries were further culled by removing (v, w) pairs whose maximum Dice Coefficient was lower than the θ threshold, for which we used the value 0.5. Finally, each word which had a mapping from the target was sequentially mapped to a majority POS tag. For example, the word *like* which had been assigned four different POS tags, IN, NN, RB, VB, by the C&C tagger in the NET Bible was only mapped to IN since the pairings of the two occurred the most frequently. The result is a mapping from one or more target lexemes to a source lexeme to a majority POS tag. In the case of *like*, two words from the target, *as* and *lijk*, were mapped thereto and to the majority tag IN.

Later, we will refer to the Wycliffe text (partially) labeled with tags projected using the Dice coefficient as DICE_1TO1.

3.1.2 GIZA++ alignments

Giza++ (Och and Ney, 2003) was also used to derive 1-to-n word alignments between the NET Bible and the Wycliffe Bible. This produces a tagged version of the Wycliffe text which we will refer to as GIZA_1TON. In our alignment experiment, we used a combination of IBM Model 1, Model 3, Model 4, and an HMM model in configuring Giza++.

GIZA_1TON was further processed to remove noise from the transferred tag set by creating a 1-to-1 word alignment: each word in the target Middle English text was given its majority tag based on the assignment of tags to GIZA_1TON as a whole. We call this version of the tagged Wycliffe text GIZA_1TO1.

3.2 Bigram tagger

Note that because the projected tags in the Wycliffe materials produced from the alignments are incomplete, there are words in the target text which have no tag. Nonetheless, a bigram tagger can be trained from maximum likelihood estimates for the words and tag sequences which were successfully projected. This serves two functions: (1) it creates a useable bigram tagger and (2) the bigram tagger can be used to fill in the gaps so that the more powerful C&C tagger can be trained on the target text.

A bigram tagger selects the most likely tag sequence T for a word sequence W by:

$$\arg\max_T P(T|W) = P(W|T)P(T)$$

Computing these terms requires knowing the transition probabilities $P(t_i|t_{i-1})$ and the emission probabilities $P(w_i|t_i)$. We use straightforward maximum likelihood estimates from data with projected tags:

$$P(t_i|t_{i-1}) = \frac{f(t_{i-1}, t_i)}{f(t_{i-1})}$$
$$P(w_i|t_i) = \frac{f(w_i, t_i)}{f(t_i)}$$

Estimates for unseen events were obtained through add-one smoothing.

In order to diversify the maximum likelihood estimates and provide robustness against the errors of any one alignment method, we concatenate several tagged versions of the Wycliffe Bible with tags projected from each of our methods (DICE_1TO1, GIZA_1TON, and GIZA_1TO1) and the NET Bible (and its tags from the C&C tagger).

3.3 Training C&C on projected tags

The bigram tagger learned from the aligned text has very limited context and cannot use rich features such as prefixes and suffixes of words in making its predictions. In contrast, the C&C tagger, which is based on that of Ratnaparkhi (1996), utilizes a wide range of features and a larger contextual window including the previous two tags and the two previous and two following words. However, the C&C tagger cannot train on texts which are not fully tagged for POS, so we use the bigram tagger to produce a completely labeled version of the Wycliffe text and train the C&C tagger on this material. The idea is that even though it is training on imperfect material, it will actually be able to correct many errors by virtue of its greater discriminitive power.

Model	Evaluate on PPCME Wycliffe		Evaluate on PPCME Test	
	PTB	COARSE	PTB	COARSE
(a) Baseline, tag NN	9.0	17.7	12.6	20.1
(b) C&C, trained on gold WSJ	56.2	63.4	56.2	62.3
(c) Bigram, trained on DICE_1TO1 and GIZA_1TON	68.0	73.1	43.9	49.8
(d) Bigram, trained on DICE_1TO1 and GIZA_1TO1	74.8	80.5	58.0	63.9
(e) C&C, trained on BOOTSTRAP (920k words)	78.8	84.1	61.3	67.8
(f) C&C, trained on BOOTSTRAP and WSJ and NET	79.5	84.8	61.9	68.5
(g) C&C, trained on (gold) PPCME Wycliffe (25k words)	n/a	n/a	71.0	76.0
(h) C&C, trained on (gold) PPCME training set (327k words)	95.9	96.9	93.7	95.1

Figure 2: Tagging results. See section 4 for discussion.

We will refer to the version of the Wycliffe text (fully) tagged in this way as BOOTSTRAP.

4 Experiments

The M3 and M34 subsections[6] of the Penn Helsinki corpus were chosen for testing since it is not only from the same period as the Wycliffe Bible but since it also includes portions of the Wycliffe Bible. A training set of 14 texts comprising 330,000 words was selected to train the C&C tagger and test the cost necessary to equal or exceed the automatic implementation. The test set consists of 4 texts with 110,000 words. The sample Wycliffe Bible with the gold standard tags has some 25,000 words.

The results of the various configurations are given in Figure 2, and are discussed in detail below.

4.1 Baselines

We provide two baselines. The first is the result of giving every word the common tag NN. The second baseline was established by directly applying the C&C tagger, trained on the Penn Treebank, to the PPCME data. The results are given in lines (a) and (b) of Figure 2 for the first and second baselines, respectively. As can be seen, the use of the Modern English tagger already provides a strong starting point for both evaluation sets.

[6]Composition dates and manuscript dates for M3 are 1350-1420. The composition dates for M34 are the same but the manuscripts date 1420-1500

4.2 Bigram taggers

In section 3.1, we discuss three versions of the Wycliffe target text labeled with tags projected across alignments from the NET Bible. The most straightforward of these were DICE_1TO1 and GIZA_1TON which directly use the alignments from the methods. Training a bigram tagger on these two sources leads to a large improvement over the C&C baseline on the PPCME Wycliffe sentences, as can be seen by comparing line (c) to line (b) in Figure 2. However, performance drops on the PPCME Test sentences, which come from different domains than the bigram tagger's automatically produced Wycliffe training material. This difference is likely to do good estimates of $P(w_i|t_i)$, but poor estimates of $P(t_i|t_{i-1})$ due to the noise introduced in GIZA_1TON.

More conservative tags projection is thus likely to have a large effect on the out-of-domain performance of the learned taggers. To test this, we trained a bigram tagger on DICE_1TO1 and the more conservative GIZA_1TO1 projection. This produces further gains for the PPCME Wycliffe, and enormous improvements on the PPCME Test data (see line (d) of Figure 2). This result confirms that conservativity beats wild guessing (at the risk of reduced coverage) for bootstrapping taggers in this way. This is very much in line with the methodology of Yarowksy and Ngai (2001), who project a small number of tags out of all those predicted by alignment. They achieve this restriction by directly adjusting the probabality mass assigned to projected tags; we do it by using two versions of the target text with tags projected in two different 1-to-1 ways.

4.3 Bootstrapping the C&C tagger

As described in section 3.3, a bigram tagger trained on DICE_1TO1 and GIZA_1TO1 (i.e., the tagger of line (d)), was used to relabel the entire Wycliffe target text to produce training material for C&C, which we call BOOTSTRAP. The intention is to see whether the more powerful tagger can bootstrap off imperfect tags and take advantage of its richer features to produce a more accurate tagger. As can be seen in row (e) of Figure 2, it provides a 3-4% gain across the board over the bigram tagger which produced its training material (row (d)).

We also considered whether using all available (non-PPCME) training material would improve tagging accuracy by training C&C on BOOTSTRAP, the Modern English Wall Street Journal (from the Penn Treebank), and the automatically tagged NET text[7] It did produce slight gains on both test sets over C&C trained on BOOTSTRAP alone. This is likely due to picking up some words that survived unchanged to the Modern English. Of course, the utility of modern material used directly in this manner will likely vary a great deal depending on the distance between the two language variants. What is perhaps most interesting is that adding the modern material did not *hurt* performance.

4.4 Upperbounds

It is apparent from the results that there is a strong domain effect on the performance of both the bigram and C&C taggers which have been trained on automatically projected tags. There is thus a question of how well we could ever hope to perform on PPCME Test given perfect tags from the Wycliffe texts. To test this, C&C was trained on the *PPCME* version of Wycliffe, which has human annotated standard tags, and then applied on the PPCME test set. We also compare this to training on PPCME texts which are similar to those in PPCME Test.

The results, given in lines (g) and (h) of Figure 2, indicate that there is a likely performance cap on non-Biblical texts when bootstrapping from parallel Biblical texts. The results in line (h) also show that the non-Biblical texts are more difficult, even with gold training material. This is likely due to the wide variety of authors and genres contained in these texts – in a sense, everything is slightly out-of-domain.

4.5 Learning curves with manual annotation

The upperbounds raise two questions. One is whether the performance gap between (g) and (h) in Figure 2 on PPCME Test is influenced by the significant difference in the size of their training sets. The other is how much gold-standard PPCME training material would be needed to match the performance of our best bootstrapped tagger (line (f)). This is a natural question to ask, as it hits at the heart of the utility of our essentially unsupervised approach versus annotating target texts manually.

To examine the cost of manually annotating the target language as compared to our unsupervised method, the C&C tagger was also trained on randomly selected sets of sentences from PPCME (disjoint from PPCME Test). Accuracy was measured on PPCME Wycliffe and Test for a range of training set sizes, sampled at exponentially increasing values (25, 50, 100, . . . , 12800). Though we trained on and predicted the full tagset used by the PPCME, it was evaluated on PTB to give an accurate comparison.[8]

The learning curves on both test sets are shown in Figure 3. The accuracy of the C&C tagger increases rapidly, and the accuracy exceeds our automated method on PPCME Test with just 50 labeled sentences and on the PPCME Wycliffe with 400 examples. This shows the domain of the target text is served much better with the projection approach.

To see how much gold-standard PPCME Wycliffe material is necessary to beat our best bootstrapped tagger, we trained the tagger as in (g) of Figure 2 with varying amounts of material. Roughly 600 labeled sentences were required to beat the performance of 61.9%/68.5% (line (f), on both metrics).

These learning curves suggest that when the domain for which one wishes to produce a tagger is significantly different from the aligned text one has available (in this and in many cases, the Bible), then labeling a small number of examples by hand is a quite reasonable approach (provided random sampling is used). However, if one is not careful, considerable effort could be put into labeling sentences

[7] This essentially is partial self-training since C&C trained on WSJ was used to produce the NET tags.

[8] Evaluation with the full PPCME set produces accuracy figures about 1% lower.

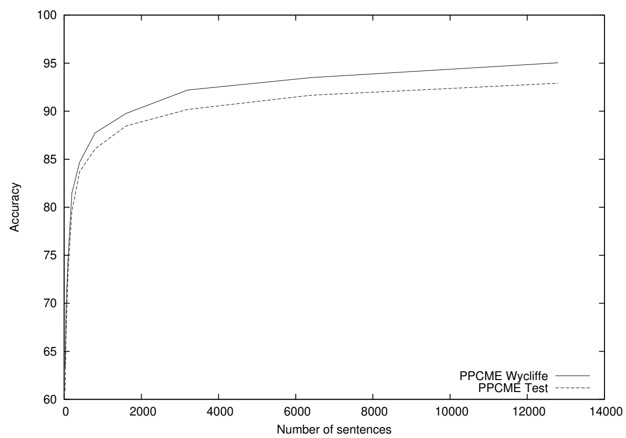

Figure 3: Learning curve showing the accuracy for PTB tags of the C&C tagger on both Bible and Test as it is given more gold-standard PPCME training sentences.

that are not optimal overall (imagine getting unlucky and starting out by manually annotating primarily Wycliffe sentences). The automated methods we present here start producing good taggers immediately, and there is much room for improving them further. Additionally, they could be used to aid manual annotation by proposing high-confidence labels even before any annotation has begun.

5 Related work

Despite the fact that the Bible has been translated into many languages and that it constitutes a solid source for studies in NLP with a concentration on machine translation or parallel text processing, the number of studies involving the Bible is fairly limited. A near exhaustive list is Chew et al.(2006), Melamed(1998), Resnik et al.(1999), and Yarowsky et al.(2001).

Yarowsky and Ngai (2001) is of central relevance to this study. The study describes an unsupervised method for inducing a monolingual POS tagger, base noun-phrase bracketer, named-entity tagger and morphological analyzers from training based on parallel texts, among many of which the Bible was included. This is particularly useful given that no manually annotated data is necessary in the target language and that it works for two languages from different families such as French and Chinese. In the case of POS tagging, only the results for English-French are given and an accuracy of 96% is achieved. Even though this accuracy figure is based on a reduced tag set smaller than the COARSE used in this study, it is still a significant increase over that achieved here. However, their method had the advantage of working in a domain that overlaps with the training data for their POS tagger. Second, the the French tag set utilized in that study is considerably smaller than the Penn Helsinki tag set, a possible source of greater noise due to its size.

Drábek and Yarowsky (2005) create a fine-grained tagger for Czech and French by enriching the tagset for parallel English text with additional morphological information, which, though not directly attested by the impoverished English morphological system (e.g. number on adjectives), typically does appear in other languages.

6 Conclusion

The purpose of the study was to implement a POS tagger for diachronic texts of maximal accuracy with minimal cost in terms of labor, regardless of the shortcuts taken. Such taggers are the building blocks in the design of higher level tools which depend on POS data such as morphological analyzers and parsers, all of which are certain to contribute to diachronic language studies and genetic studies of language change.

We showed that using two conservative methods for projecting tags through alignment significantly improves bigram POS tagging accuracies over a baseline of applying a Modern English tagger to Middle English text. Results were improved further by training a more powerful maximum entropy tagger on the predictions of the bootstrapped bigram tagger, and we observed a further, small boost by using Modern English tagged material in addition to the projected tags when training the maximum entropy tagger.

Nonetheless, our results show that there is still much room for improvement. A manually annotated training set of 400–800 sentences surpassed our best bootstrapped tagger. However, it should be noted that the learning curve approach was based on domain neutral, fully randomized, incremental texts, which are not easily replicated in real world applications. The domain effect is particularly evident in training on the sample Wycliffe and tagging on the test PPCME set. Of course, our approach can be integrated with one based on annotation by using our bootstrapped taggers to perform semi-automated annotation, even *before* the first human-annotated tag has been labeled.

It is not certain how our method would fare on the far more numerous parallel diachronic texts which do not come prealigned. It is also questionable whether it would still be robust on texts predating Middle English, which might as well be written in a foreign language when compared to Modern English. These are all limitations that need to be explored in the future.

Immediate improvements can be sought for the algorithms themselves. By restricting the mapping of words to only one POS tag in the Wycliffe Bible, this seriously handicapped the utility of a bigram tagger. It should be relatively straightforward to transfer the probability mass of multiple POS tags in a modern text to corresponding words in a diachronic text and include this modified probability in the bigram tagger. When further augmented for automatic parameter adjustment with the forward-backward algorithm, accuracy rates might increase further. Furthermore, different algorithms might be better able to take advantage of similarities in orthography and syntactic structure when constructing word alignment tables. Minimum Edit Distance algorithms seem particularly promising in this regard.

Finally, it is evident that the utility of the Bible as a potential resource of parallel texts has largely gone untapped in NLP research. Considering that it has probably been translated into more languages than any other single text, and that this richness of parallelism holds not only for synchrony but diachrony, its usefulness would apply not only to the most immediate concern of building language tools for many of the the world's underdocumented languages, but also to cross-linguistic studies of unprecedented scope at the level of language genera. This study shows that despite the fact that any two Bibles are rarely in a direct parallel relation, standard NLP methods can be applied with success.

References

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Helena Britto, Marcelo Finger, and Charlotte Galves, 2002. *Computational and linguistic aspects of the construction of The Tycho Brahe Parsed Corpus of Historical Portuguese*. Tübingen: Narr.

Peter A. Chew, Steve J. Verzi, Travis L. Bauer, and Jonathan T. McClain. 2006. Evaluation of the bible as a resource for cross-language information retrieval. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability, Sydney, July 2006*, pages 68–74.

James R Curran and Stephen Clark. 2003. Investigating gis and smoothing for maximum entropy taggers. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-03)*.

Lee R. Dice. 1945. Measures of the amount of ecologic association between species. *Journal of Ecology*, 26:297–302.

Elliott Franco Drábek and David Yarowsky. 2005. Induction of fine-grained part-of-speech taggers via classifier combination and crosslingual projection. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 49–56, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Barbara A. Fennell. 2001. *A History of English: A Sociolinguistic Approach*. Blackwell, Oxford.

Chung-Hye Han, 2000. *The Evolution of Do-Support In English Imperatives*, pages 275–295. Oxford University Press.

Martin Kay and Martin Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Anthony Kroch and Ann Taylor. 2000. Penn-helsinki parsed corpus of middle english, second edition.

Anthony Kroch, Ann Taylor, and Donald Ringe. 2000. The middle english verb-second constraint: A case study in language contact and language change. *Amsterdam Studies in the Theory and History of Linguistic Science Series*, 4:353–392.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Dan I. Melamed. 1998. Manual annotation of translation equivalence: The blinker project. In *Technical Report 98-07, Institute for Research in Cognitive Science, Philadelphia*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sylwia Ozdowska. 2006. Projecting pos tags and syntactic dependencies from english and french to polish in aligned corpora. In *EACL 2006 Workshop on Cross-Language Knowledge Induction*.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the "book of 2000 tongues". *Computers and the Humanities*, 33(1–2):129–153.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT '01: Proceedings of the first international conference on Human language technology research*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Appendix

Figure 4 provides the full mapping from PPCME tags to the Penn Treebank Tags used in our evaluation.

PPCME→PTB	PPCME→PTB
ADJR→JJR	N→NN
ADJS→JJS	N$→NN
ADV→RB	NEG→RB
ADVR→RBR	NPR→NNP
ADVS→RBS	NPR$→NNP
ALSO→RB	NPRS→NNPS
BAG→VBG	NPRS$→NNPS
BE→VB	NS→NNS
BED→VBD	NS$→NNS
BEI→VB	NUM→CD
BEN→VBN	NUM$→CD
BEP→VBZ	ONE→PRP
C→IN	ONE$→PRP$
CODE→CODE	OTHER→PRP
CONJ→CC	OTHER$→PRP
D→DT	OTHERS→PRP
DAG→VBG	OTHERS$→PRP
DAN→VBN	P→IN
DO→VB	PRO→PRP
DOD→VBD	PRO$→PRP$
DOI→VB	Q→JJ
DON→VBN	Q$→JJ
DOP→VBP	QR→RBR
E_S→E_S	QS→RBS
ELSE→RB	RP→RB
EX→EX	SUCH→RB
FOR→IN	TO→TO
FOR+TO→IN	VAG→VBG
FP→CC	VAN→VBN
FW→FW	VB→VB
HAG→VBG	VBD→VBD
HAN→VBN	VBI→VB
HV→VB	VBN→VBN
HVD→VBD	VBP→VBP
HVI→VB	WADV→WRB
HVN→VBN	WARD→WARD
HVP→VBP	WD→WDT
ID→ID	WPRO→WP
INTJ→UH	WPRO$→WP$
MAN→PRP	WQ→IN
MD→MD	X→X
MD0→MD	

Figure 4: Table of mappings from PPCME tags to Penn Treebank Tags.