Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 38 (2005) 422–430
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Domain-specific language models and lexicons for tagging

Anni R. Coden [a,*], Serguei V. Pakhomov [b], Rie K. Ando [a], Patrick H. Duffy [b], Christopher G. Chute [b]

[a] IBM, T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532, USA
[b] Division of Medical Informatics Research, Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA

Received 16 July 2004
Available online 2 April 2005

Abstract

Accurate and reliable part-of-speech tagging is useful for many Natural Language Processing (NLP) tasks that form the foundation of NLP-based approaches to information retrieval and data mining. In general, large annotated corpora are necessary to achieve desired part-of-speech tagger accuracy. We show that a large annotated general-English corpus is not sufficient for building a part-of-speech tagger model adequate for tagging documents from the medical domain. However, adding a quite small domain-specific corpus to a large general-English one boosts performance to over 92% accuracy from 87% in our studies. We also suggest a number of characteristics to quantify the similarities between a training corpus and the test data. These results give guidance for creating an appropriate corpus for building a part-of-speech tagger model that gives satisfactory accuracy results on a new domain at a relatively small cost.

Keywords: Clinical report analysis; Part-of-speech tagging accuracy; Domain adaptation; Clinical information systems; Biomedical domain; Corpus linguistics; Statistical part-of-speech tagging; Hidden Markov Model

1. Introduction

Accurate and reliable part-of-speech (POS) tagging is useful for many Natural Language Processing (NLP) tasks such as syntactic parsing, feature extraction for classification, semantic representation, and among others that, in turn, form the foundation of NLP-based approaches to information retrieval and data mining. Many high precision statistical POS taggers [1,2] are available both in the open source and proprietary domains. For research purposes, taggers are in general trained and tested on a general-purpose corpus of annotated text such as the Penn Treebank-2 corpus [7] distributed by the Linguistic Data Consortium (LDC). Whereas the accuracy of tagging such general English data is very high, it usually entails starting with a relatively large amount of training data and/or a complete lexicon. When the tagger is used for a new "sub-language" such as the medical sub-domain, one expects to find a large number of new lexical items for which a tagger trained on general English may not have sufficient statistical and other information. In statistical POS tagging, this problem is typically addressed by adapting the training data and lexicons to the target domain, which constitutes the focal point of this paper.

Our main goal in this paper is to quantify the differences between general English and a specialized sub-language domain of medical English with respect to part-of-speech assignments. Our main methodological research is to uncover the trade-offs in adapting a general-purpose statistical POS tagger to a medical English sub-domain. We examine and compare two types of adaptation. One involves manually annotating a number of documents from the target domain and adding

* Corresponding author. Fax: +1 914 784 6054.
E-mail address: anni@us.ibm.com (A.R. Coden).

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.02.009

these newly annotated documents to the general English training data. The other adds a lexicon derived from the target domain.

We discuss related work in Section 2 and present problem descriptions in Section 3. In particular, we introduce a quantitative analysis of the differences in the characteristics (e.g., part-of-speech assignments, vocabulary) as well as their distributional properties across three corpora: Treebank-2, GENIA, and MED (a manually tagged corpus of medical clinical notes). We quantify the differences among the three corpora. In Section 4, we report on a set of experiments using several combinations of the corpora for training and testing. Finally, we also report on a set of experiments with domain lexicons. We show that a model built from a small domain corpus in conjunction with a general English corpus improves the accuracy of a tagger substantially. On the other hand, a domain lexicon used together with a model built from general English corpus has only a small impact, but at a fraction of the cost of manually annotating a domain corpus.

2. Related work

POS tagging is one of the better understood and addressed problems in the NLP community. In general, state-of-the-art POS tagging technology is highly accurate. It has been shown that high accuracy can be achieved by taggers that do not use hand-crafted rules but instead rely on mathematical models such as Hidden Markov Models (HMM) (e.g., [1,4,17]), maximum entropy models [12], and transformation-based learning models [3].

These taggers automatically learn model parameters (probabilities or transformation rules) from training corpora that are manually annotated with part-of-speech tags.[1] The underlying assumption is that the test data (the data we need to process in practical applications) and the training data are drawn from the same type of discourse and, thus, share distributional characteristics. In addition, the training corpus needs be sufficiently large (typically over one million words) for the models to obtain reliable statistics. According to the literature, the different types of statistical taggers achieve essentially similar high accuracy upon the availability of such appropriate training data. For our experiments, we will use an HMM tagger as discussed in more detail in Section 3.

Achieving as high accuracy when the training corpus and the test data are part of different types of discourse is a challenge. It is difficult and expensive to develop a domain-specific training corpus and one can safely assume that the unknown word rate increases substantially when the training corpus and test data differ in type.

There are several examples in the literature on how unknown words degrade tagger accuracy. For example, evaluations of Brants's HMM-based TnT tagger with smoothing and unknown word prediction modules show an overall accuracy of 96.7% on both the NEGRA corpus of German and the Penn Treebank of general English [1]. While the TnT tagger performs at 97% accuracy on known words in the Penn Treebank corpus, the accuracy drops to 89% on unknown words. The LT POS tagger is reported to perform at 93.6–94.3% accuracy on known words and at 87.7–88.7% on unknown words using a cascading guesser [9]. The overall results for both of these taggers are much closer to the high end of the spectrum because the rate of the unknown words in the tests performed on the Penn Treebank corpus is generally relatively low—2.9% [1]. From such results, we can conclude that the higher the rate of unknown vocabulary, the lower the overall accuracy will be, necessitating the adaptation of the taggers trained on the Penn Treebank corpus to sub-language domains with vocabulary that is substantially different from the one represented by the Penn Treebank corpus.

Rindflesch et al. [13] report 93.1% accuracy achieved with the Xerox [4] tagger. The tagger is trained on MEDLINE abstracts with a medical lexicon; however, it uses a SPECIALIST lexicon annotated with fewer POS categories than the standard Penn Treebank tagset, which makes comparisons difficult without reducing the Penn Treebank tagset to the SPECIALIST tagset. Smith et al. [15] designed an HMM-based POS tagger (MedPost) and trained it on hand-annotated MEDLINE abstracts. They report over 97% accuracy on 1000 sentences from biomedical articles. Smith et al. also find that using a domain lexicon in combination with a domain corpus for training HMM-based taggers such as MedPost happen to be more beneficial than using a tagger trained purely on general English data such as the Brown corpus and the Wall Street Journal data represented in the Penn Treebank corpus (Rindflesch, p.c.).

Jensen et al. [5] report on another example of tagger adaptation to the biomedical domain. In their work on using biomedical literature for knowledge discovery, Jensen et al. report the results of re-training a TreeTagger [14] on the GENIA corpus. The tagger trained with the Treebank (the authors refer to it as the UPenn corpus) was accurate on 85.7% of the test data (manually tagged MEDLINE abstracts). Retraining it with GENIA data improved the results to 93.6%. Unfortunately, the authors do not present the details of their experiments with POS tagging. For example, it is unclear how much data were used for training and testing. However, the results indicate that domain adaptation results in improved performance.

[1] The Brill tagger has to be "seeded" with handcrafted rule templates.

3. Problem description

The problem this work addresses is how to adapt a POS tagger to the biomedical domain, given the availability of a large general English training corpus. We focus here on two medical sub-domains, one of them being clinical notes dictated by physicians in the course of seeing patients and filed as part of the patient's chart, the other being biomedical literature abstracts published in PubMed. We will explore the characteristics of these two corpora and compare them with the characteristics of the Penn Treebank-2 corpus to gather insights into appropriate models for POS tagging. We focus on characteristics typically used by statistical POS taggers.

For our experiments, we used Hidden Markov model taggers, which assume that a Markov process, whose states correspond to POS tags, emits a sequence of words. We call it an n-gram model when the current tag (i.e., the current state of the Markov process) is assumed to depend on the preceding $(n-1)$ tags. For instance, according to a bi-gram model, the probability that we observe a sentence of m words, $w_1 \dots w_m$ with tags $t_1 \dots t_m$ can be written as: $P(w_1 \dots w_m; t_1 \dots t_m) = \Pi_{i=1,m}\ P(w_i \mid t_i)\ P(t_i \mid t_{i-1})$. The tagging problem is to find the tag sequence $t_1 \dots t_m$ that maximizes the likelihood. Word emission probabilities $P(\text{word} \mid \text{tag})$ and tag-transition probabilities $P(\text{tag} \mid \text{preceding-tags})$ are estimated from frequencies observed in the pre-annotated training data. The estimation $P(\text{word} \mid \text{tag})$ can be approximated by frequency(word, tag)/frequency(tag) which is known to be poor on low-frequency words. However, the words that do not appear in the training data (out-of-vocabulary words) would obtain zero probability, which would make the above likelihood maximization incomputable. There are a number of smoothing techniques to address this problem [18]. In our experiments, as is commonly practiced, we use character types, endings of words, and tag distributions over low-frequency words for predicting parts-of-speech of unseen (or rarely observed) words. More precisely, we estimate the probability $P(\text{word} \mid \text{tag})$ for rare words (with frequency <5 in the training data) by: $f(\text{rare\_word} \mid \text{tag})$, $f(\text{char-type} \mid \text{tag})$ and $\sum_l \lambda_l f(\text{ending}_l \mid \text{tag})$, where ending_l denotes endings of length l, and where $f(x \mid y) = \text{frequency}(x,y)/\text{frequency}(y)$ and coefficients λ_l are determined using standard deviation as in [1]. There are several formulations of smoothing for POS tagging that are known to be equivalently effective. The specific choice is not central to the theme of this paper.

We tested four types of models: uni-, bi-, and tri-gram models with smoothing as described above, and a unigram model with simplified smoothing—hereafter, abbreviated as 'uni-gram,' 'bi-gram,' 'tri-gram,' and 'uni-gram-no,' respectively. Bi-gram and tri-gram models are widely used in practical settings. The motivation for using a uni-gram model (which does not rely on tag-transition probabilities) is to see how the absence of tag-transition statistics affects performance. Uni-gram-no is the simplest model, which does not use tag-transition statistics and performs minimum smoothing. That is, it determines the most frequent POS tag in the training corpus and assigns this tag to all the out-of-vocabulary words.

The performance differences between using a unigram and uni-gram-no model will show how much sophisticated smoothing helps to counteract the out-of-vocabulary word problem. Our interest is, especially, in settings where the tag-transition statistics of the test data may be quite different from those of the training data. Also, note that the proportion of out-of-vocabulary words may be high when the training and test corpora are from different domains.

3.1. The corpora

Our experiments involve three corpora, the Penn Treebank-2 [8] corpus, the GENIA corpus [6], and the MED corpus of clinical notes. In particular, we use a subset (hereafter TB-2) of the Penn Treebank corpus that consists of the Brown and Wall Street Journal collections distributed by the LDC, a large corpus, that has been manually annotated with POS tags, and is widely used to train taggers.

The GENIA corpus [6] is a set of 2000 Medline abstracts obtained by using three different search key words: "Human," "Blood Cells," and "Transcription Factors." This corpus has also been manually annotated with POS tags [16]. However, the annotation guidelines differ slightly from those for TB-2 and MED. In particular, proper noun tags are not used in annotating the GENIA corpus except for bibliographical information (e.g., authors, research institutes) and the tag representing special symbols was intentionally used sparsely.

A proprietary MED corpus was developed at the Mayo Clinic in Rochester, Minnesota. It is the goal of this medical institution to tag their ever-growing set of clinical notes with POS information. The Clinical notes repository at the Mayo Clinic consists of all documents dictated by physicians and subsequently transcribed and filed as part of the patients' electronic medical record (EMR). The notes follow the HL7 Clinical Document Architecture standard [19] where the information is templated into sections such as Chief Complaint, Current Medications, and Impression/Plan among others. The repository contains outpatient notes as well as discharge summaries and inpatient service notes. The collection does not contain the inpatient progress notes, however. Due to the fact that the notes are initially dictated via a speech interface, they represent quasi-spontaneous discourse [10], which is characterized by phenomena typically found in spontaneous speech such as disfluencies,

ellipses, ungrammatical sentences, spelling, and punctuation errors. All of these factors contribute to increased difficulty in processing this corpus for POS information.

The current size of the MED collection is approximately 16 million documents. It is growing at the rate of 40,000–60,000 documents per week. To create a clinical notes corpus for POS tagging, 273 clinical notes were picked randomly from the pool of clinical notes and manually annotated with POS tags. Three domain experts familiar with the language of the clinical notes annotated the collection. Pakhomov et al. [11] showed that the inter-annotator agreement is reliable at $\kappa = 0.93$. The following is a passage from a typical clinical note:

MED: *# 1 Left ACL disruption, return-to-work evaluation Patient of Dr. NAME. Samples mailed to home address. Patient is on Prilosec 20 mg bid. The ACE level remains in the lower limit of normal. Total cholesterol is 160 with an HDL cholesterol of 43, and LDL of 92, and a triglyceride of 123.*

In contrast, a sentence from a Medline article from the GENIA collection and a sentence from the Penn Tree Bank corpus are shown next.

GENIA: *TI - IL-2 gene expression and NF-jB activation through CD28 requires reactive oxygen production by 5-lipoxygenase. AB-activation of the CD28 surface receptor provides a major costimulatory signal for T-cell activation resulting in enhanced production of interleukin-2 (IL-2) and cell proliferation.*

Penn TreeBank: *The asbestos fiber, crocidolite, is unusually resilient once it enters the lungs, with even brief exposures to it causing symptoms that show up decades later, researchers said. Lorillard Inc., the unit of New York-based Loews that makes Kent cigarettes, stopped using crocidolite in its Micronite cigarette filters in 1956.*

We will qualify and quantify similarities and differences among these three corpora in the subsequent section.

3.2. Similarities and differences of corpora

We present the statistics pertaining to the three corpora in terms of 'words' and 'word types.' A corpus is split into units, each individual instance of such a unit is a *word*. Words are converted to all lower case. All identical words (i.e., having exactly the same spelling) belong to the same *word type*. Table 1 shows the number of words and word types in the three corpora.

The POS tagger used in our study applies a common number normalization algorithm: each occurrence of a digit in a word is mapped to the digit 0. For example, the number 3 is mapped to 0, 55 is mapped to 00, and L8 is mapped into L0. The underlying assumption is that words that differ only in digits are essentially the same from the perspective of tagging. Table 2 shows the percentage decrease in word types due to number normalization. The biggest drop is seen in the GENIA corpus indicating that many words differ only in digits. For example, "# -fold" where # is a one or more digits, appears frequently.

Table 1
Words and word type counts for TB-2, MED, and GENIA corpora

	TB-2	MED	GENIA
# words	1,289,212	100,650	501,062
# word types	45,684	8,702	22,534

To compare the vocabulary sizes of the three corpora, we counted the number of word types in the first 100,000 words in each corpus. The results are shown in Table 3.

It is not surprising that the GENIA collection has the smallest vocabulary, as its documents are the results of a three keyword query. The MED corpus content is limited to clinical diagnoses, observations, and other clinically relevant topics. Some sections within clinical notes such as social history and family history tend to have a broader coverage; however, the vocabulary still revolves around a limited number of topics. On the other hand, TB-2 covers a relatively wide range of topics, thus it is not surprising that its vocabulary size is greater that that of the MED corpus.

Table 4 shows the average sentence length in the three corpora. These numbers indicate that the MED corpus consists of much shorter sentences than the other corpora.

Table 2
Percentage decrease of word types due to number normalization for TB-2, MED, and GENIA corpora

	TB-2	MED	GENIA
% decrease of word types	13.30	8.70	19.53

Table 3
Vocabulary size for 100 K words of TB-2, MED, and GENIA corpora

Corpus	# words	# word types
TB-2	100,000	10,576
MED	100,000	7,937
GENIA	100,000	7,410

Table 4
Number of sentences and average sentence length for TB-2, MED, and GENIA corpora

	# words	# sentences	Average sentence length
TB-2	1,289,212	53,362	24.16
MED	100,650	7,299	13.79
GENIA	501,062	18,436	27.18

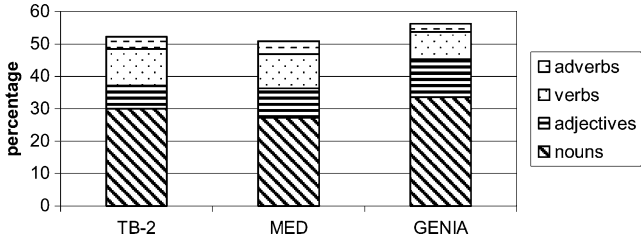

Fig. 1. Comparisons of POS tag distributions for most frequent tags in TB-2, MED, and GENIA corpora. The rest of the tags are for function words, punctuations, and numbers.

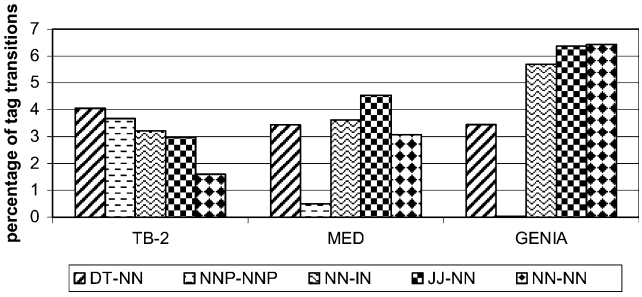

Fig. 2. Comparisons of five most frequent transitions between POS tags in TB-2, MED, and GENIA corpora.

A portion of the MED corpus consists of sentence fragments. For instance, some are missing the explicit mention of the subject when the subject of the sentence is about the patient. "Winters in Florida" is an example of such a sentence fragment. Another frequent type of sentence fragment consists of the transcription of spoken attribute-value pairs recorded during physical examination such as "Lungs – clear. Throat – erythematous. BP: 120/80." Apart from fragments, the fact that clinical notes represent a transcript of quasi-spontaneous spoken discourse may contribute to shorter sentence length, since we tend to communicate in smaller chunks in speech than in text.

Shorter sentences may present a challenge to POS tagging, which relies on context for correct classification by limiting the context for some words that happen to be consistently close to sentence boundaries. For example, a human can deduce easily from world knowledge that "Winters" in "Winters in Florida" is a verb, however an automatic POS tagger may have problems correctly tagging this word because of the limited available linguistic context.

Fig. 1 shows the tag distribution by tag groups. Lexical variations of nouns, adjectives, verbs, and adverbs are grouped together.

GENIA has a higher percentage of nouns and adjectives and a lower percentage of verbs. The distributions of tags in TB-2 and MED corpora are quite similar.

As described in the beginning of Section 3, taggers use transition statistics. Fig. 2 shows the five most frequent transitions in the three corpora as a percentage of all transitions. The following abbreviations are used: DT for *determiners*, NN for *nouns*, NNP for *proper nouns*, IN for *prepositions* or *subordinating conjunctions*, and JJ for *adjectives*.

The transition statistics (Fig. 2) in conjunction with other corpora statistics lead to some more observations. The transition between determiners and nouns is higher in TB-2 than the other corpora. However, in fact, the percentage of words classified as determiners is nearly the same in all three corpora. This is attributable to the fact that both MED and GENIA corpora have a larger proportion of noun phrases with nominal (NN–NN) and adjectival modification (JJ–NN) than the TB-2 corpus. The proportion of both nouns and determiners is roughly the same across all three corpora, but there is a higher proportion of NN–NN and JJ–NN transitions in the MED and GENIA corpora, it is reasonable to conclude that nominal compounds and adjectival modifiers are responsible for the reduction in the proportion of DT–NN transitions. There are hardly any proper nouns tagged in the GENIA corpus, which explains why there are no proper noun transitions among the top five transitions for that corpus.

4. Adaptation study

In this section, we study POS tagging performance, with the goal of improving accuracy on medical domain corpora. In the first suite of experiments, the training and test data are from the same domain (Section 4.1). The results serve as the baseline (reference performance) for our succeeding experiments. Next, we demonstrate that POS tagging performance significantly degrades when TB-2—the standard corpus—is used as training data for tagging the medical corpora (Section 4.2). We explore two types of training procedures for improving on medical corpora in Sections 4.3 and 4.4.

The following experimental framework is used for our evaluation. As described in Section 3, we test four models: uni-gram-no (uni-gram model, no unknown word type processing), uni-gram, bi-gram, and tri-gram (all with unknown word type processing). When a corpus needs to be split into the training set and test set, we made 10 runs—each of which uses randomly chosen 80% of the corpus as training data and the rest as test data—, and report the average performance over these 10 runs. Our evaluation metric is accuracy: # (correctly tagged words)/# (all words).

4.1. Training and test corpus from same domain

To establish the first baseline, the training corpus and the test corpus are derived from the same domain. The

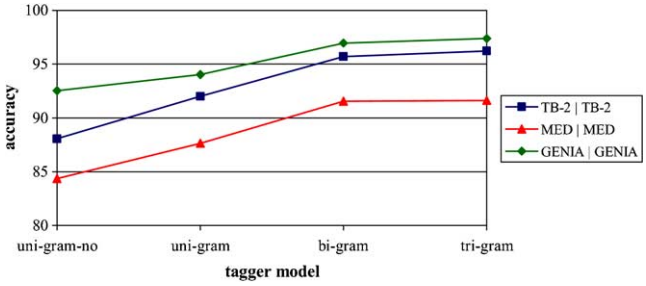

Fig. 3. Baseline POS tagging accuracy, training, and testing data derived from TB-2, MED or GENIA corpus, respectively.

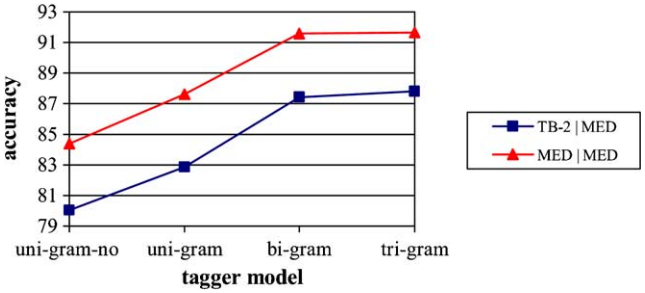

Fig. 4. POS tagging accuracy: training data TB-2 or MED, test data MED.

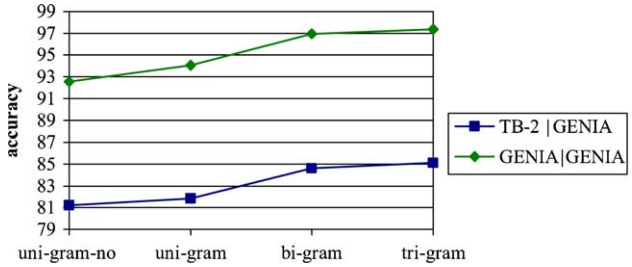

Fig. 5. POS tagging accuracy: training data TB-2 or GENIA, test data GENIA.

results of these runs are shown in Fig. 3. The labels indicate the training and test set. For instance "TB-2 | TB-2" means that the training and the test sets were derived from the TB-2 corpus. In cases when these two sets are the same, the accuracy numbers reported are based on our experimental framework as described earlier in this section.

The TB-2 and GENIA runs show good accuracy, whereas the accuracy on the MED data needs to be improved for practical applications. Low accuracy is caused by the relatively small size of the training corpus. Recall, that MED contains only 100,650 words whereas TB-2 is 1,289,212 words (Table 1). Our goal is to improve performance without manually tagging a bigger training corpus.

Another factor we examined is the percentage of unambiguous word types. '*Unambigous*' is defined here as having a single tag associated with a word type within a single corpus. In particular, in the GENIA corpus 90.74% of the word types are unambiguous, whereas TB-2 and MED have only 83.62 and 83.84% unambiguous word types, respectively. Hence, tagging the GENIA is easier than tagging MED or TB-2.

It is surprising at first to see how well uni-gram-no performs on the GENIA corpus. Examining the average out-of-vocabulary rates sheds some light on these results as shown in Table 5.

MED corpus has a very high out-of-vocabulary rate in comparison to the other corpora and also the largest standard deviation.

4.2. Training with TB-2

For practical use, it is desirable to have a higher accuracy than the 92% produced by tri-gram models on MED data. Towards this end, we trained the POS tagger with the large TB-2 corpus and tested it on MED and GENIA. The results are given in Fig. 4 and Fig. 5.

Not surprisingly, such an approach does not work very well with the MED corpus, in spite of the small size of the corpus. We show in Fig. 4, that the accuracy degrades substantially when trained with TB-2 as compared to when trained with the (small) MED corpus. The out-of-vocabulary rate, which is 10.18% when trained with 80% of the MED corpus, increases to 12.47% when trained with the TB-2 corpus. Other reasons for such degradation are the differences in tag distributions and tag-transition distributions as described in Section 3.2.

The accuracy degrades even more dramatically on GENIA as shown in Fig. 5.

The out-of-vocabulary rate is 4.32% when trained with GENIA, and it increases to an average of 21.24% when trained with TB-2.

Table 5
Out-of–vocabulary rate for TB-2, MED, and GENIA corpora

	TB-2	MED	GENIA
% OOV	3.66	10.18	4.32
SD	0.038	0.137	0.065

4.3. Adaptation with domain corpus

We have observed that a general English corpus like TB-2 is not sufficient as a training corpus in the medical domain. The question arises whether performance can be improved by adding some medical corpus to the TB-2 corpus for training purposes. We measured the accuracy of the tagger trained on the TB-2 and GENIA (or MED) corpus and tested on the MED (or GENIA) corpus. Again, our motivation is to save the high cost of developing a large domain-specific training corpus. It would be desirable if, for instance, the publicly available

GENIA corpus could be used to enhance tagging accuracy on different types of medical corpora like the MED corpus. Although, adding GENIA to the TB-2 corpus improves the performance on MED slightly for some tagger models, the improvement is quite small as can be seen in Fig. 6.

Adding MED to the TB-2 corpus for training does not change the performance on GENIA. These results are not surprising, as the GENIA corpus and MED corpus share only a few word types. In particular, Fig. 7 shows the number of distinct word types in each of the three corpora and their mutual overlap. Only 2418 word types are present in all three corpora. Adding the GENIA corpus to the TB-2 corpus for training should not improve the performance when tested on MED as only 593 new word types also present in MED are added to the training corpus.

It seems that a training corpus similar to the testing corpus is necessary to boost tagger accuracy from levels achieved when training with only a general English corpus. Is a small additional training corpus sufficient? Fig. 8 compares three different set of runs based on different training corpora.

Tagger accuracy improves more when adding a domain-specific corpus to the training corpus. Improvements are particularly large with uni-gram-no model.

However, adding the GENIA corpus to the TB-2 corpus for training changes the accuracy minimally as depicted in Fig. 9.

This is not surprising as the out-of-vocabulary rate is only 4.32% when the training data are drawn from 80% of the GENIA corpus and test data are the remaining 20% as shown in Table 5. When the training corpus consists of TB-2 and 80% of the GENIA corpus, the out-of-vocabulary rate drops to approximately 2%. Hence, we see only a very slight performance improvement.

In general, one would assume that a bigger training corpus would boost the accuracy by reducing the out-of-vocabulary rate. However, adding a corpus can also decrease the accuracy. A tri-gram model trained with TB-2 only achieves an accuracy of 85.1% on GENIA. When the MED corpus is added to the TB-2 training set, the accuracy drops to 84.41%. A contributing factor is the differences in tag sets associated with a word type between two corpora. In Section 4.1, we introduced the notion of an ambiguous word type: a word type is ambiguous if it has multiple tags associated with it within a corpus. Clearly, a higher percentage of ambiguous word types increases the difficulty of POS tagging. Combining two corpora for training, could increase this per-

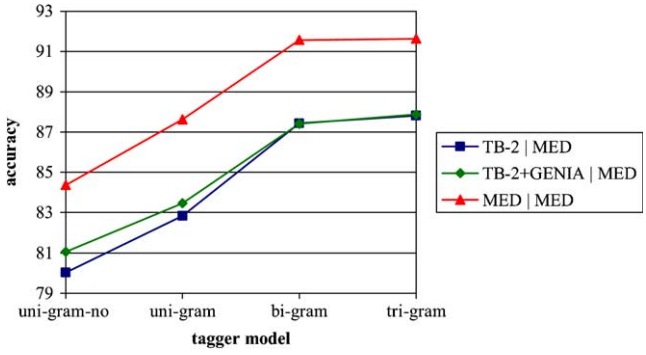

Fig. 6. POS tagging accuracy: training data TB-2, TB-2 + GENIA, or MED, test data MED.

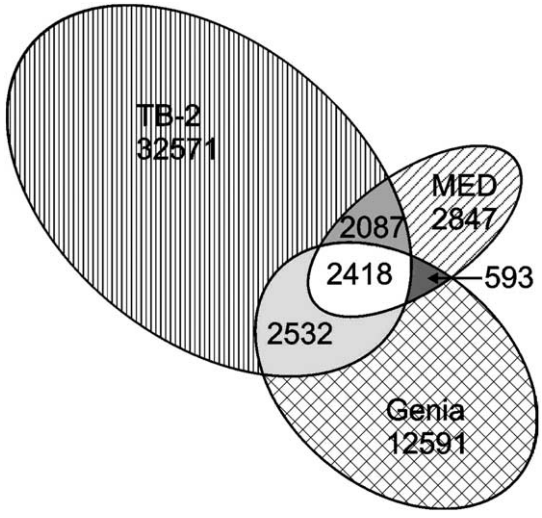

Fig. 7. Overlap of word types between TB-2, MED, and GENIA corpora.

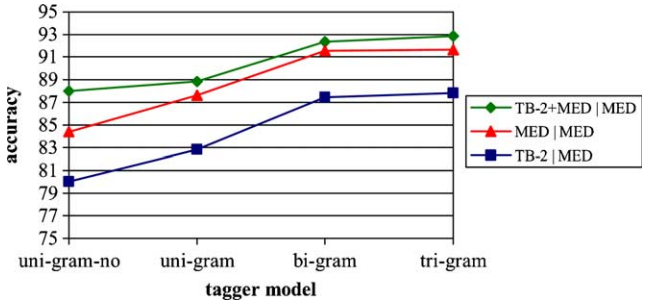

Fig. 8. Comparison of POS tagging accuracy with respect to different training corpora: MED, TB-2 + MED, tested on MED.

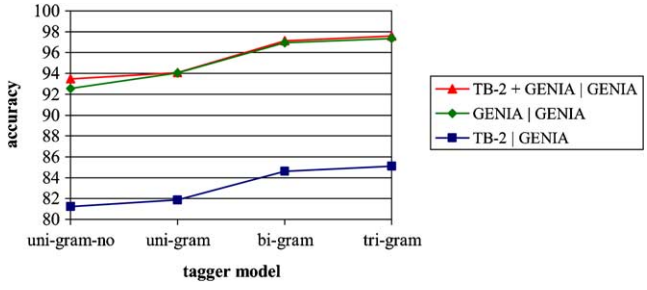

Fig. 9. Comparison of POS tagging accuracy with respect to different training corpora: TB-2, MED, TB-2 + MED, tested on GENIA.

Table 6
Comparison of POS tagging accuracy: training data TB-2 augmented with lexicons of various sizes, tested on MED

	TB-2	+L100	+L200	+L300	+L400	+L500
uni-gram-no	80.03	81.35	81.71	82.07	82.23	82.48
uni-gram	82.85	83.51	83.63	83.77	83.78	83.87
bi-gram	87.44	88.08	88.26	88.39	88.39	88.46
tri-gram	87.82	88.42	88.58	88.72	88.74	88.82

centage. For example, ''implant'' is tagged as a verb ''VB'' in the TB-2 corpus, whereas it is tagged as noun ''NN'' in the MED corpus. The word ''sore'' is tagged only as an adjective ''JJ'' in the TB-2 corpus, whereas in the MED corpus it is tagged both as an adjective ''JJ'' and as a noun ''NN.'' In fact, approximately half of the word types common to TB-2 and MED have different tag sets depending to which corpus they belonged. Such characteristics contribute to the reduction in accuracy in the above mentioned experiment of adapting a training corpus to a different domain.

4.4. Use of lexicon

Can a domain lexicon be used instead of adding a domain training corpus? We computed the 500 most frequent word types from the pool of 16 million clinical notes collection and removed some word types: stop words as well as abbreviations indicating section headings were not included despite their high frequency. The word types in the lexicon were manually POS tagged. It is noteworthy, that 482 out of the 500 word types in the lexicon were in the MED corpus. This indicates that the vocabulary in the sampling of clinical notes in the MED corpus is representative of the general collection. To use lexicons for building models, we 'pretend' that each of the word-tag pairs in the lexicon occurred just once in the training set. Thus, a lexicon affects the estimations of word emission probabilities.

We built five lexicons of 100–500 word types and used each of them in conjunction with a model trained with TB-2 to tag the MED corpus. Even a small lexicon improves the performance over using the model without any domain knowledge as shown in Table 6.

The accuracy improvement grows with the size of the lexicon. We showed previously that a model trained with 80% of MED and TB-2 yields an accuracy of 92.87% with a tri-gram model. A model trained on TB-2 and a lexicon of 500 word types yields an accuracy of 88.82%. Hence, training with a domain model and TB-2 yields a 4.56% improvement over training with a lexicon and TB-2. The question arises whether a model trained with 80% of MED, TB-2, and a lexicon on 500 words would yield accuracies in excess of 92.87% with a tri-gram model. Our experiments showed an accuracy of 92.88% in that case, a quite insignificant improvement. This is not surprising, as there is a high overlap of word types in the lexicon and word types in the MED corpus as previously stated.

5. Conclusion

POS tagging forms a basis for many different natural language applications. Smith [15] observes that ''a 4% error rate corresponds approximately to one error per sentence'' necessitating a high accuracy. We showed that a tagger using a general-purpose English model, like one build from the TB-2 corpus, does not perform satisfactory when tagging medical discourse like clinical notes or PubMed abstracts. We show that training with a small domain-specific corpus (e.g., MED) in addition to a general-English corpus (e.g., TB-2) boosts the accuracy by 5.75–9.94% over tagging with a general-English training corpus only. Domain lexicons used in conjunction with a general-English training corpus also boost the accuracy (although not as much) and are much cheaper to develop.

We furthermore analyzed the characteristics of three corpora, TB-2, GENIA, and MED to quantify why a tagger model using one of the corpora is not necessarily adequate to POS tag a different corpus. Our studies showed that our HMM tagger can achieve 92% accuracy when its model is built from a general-English corpus in conjunction with a small domain corpus. To achieve the same accuracy on the GENIA corpus, the model has to be built from (part of) the GENIA corpus. Adding a general-English corpus to build the model does not change the accuracy of the tagger. However, using a domain corpus (i.e., GENIA) accuracy of 97% can be achieved. It remains to be seen whether the performance of the tagger using a general English model and a sufficiently large domain lexicon has the same accuracy as training with a domain corpus.

References

[1] Brants T, TnT–A statistical part-of-speech tagger. In: Proceedings of the sixth applied natural language processing conference (ANLP-2000). p. 224–31.
[2] Brill E, A corpus-based approach to language learning. Ph.D. Dissertation, Department of Computer and Information Science, University of Pennsylvania; 1993.

[3] Brill E, Some advances in rule-based part of speech tagging. In: Proceedings of the 12th national conference on artificial intelligence 1994 (AAAI-94). p. 722–7.
[4] Cutting D, Kupiec J, Pedersen J, Sibun P. A practical part-of-speech tagger. In: Proceedings of the third conference on applied natural language processing 1992 (ANLP-92), Trento, Italy; 1992.
[5] Jensen L, Saric J, Bork P. Utilizing literature for biological discovery. In: Proceedings of E-BioSci/ORIEL. Varenna, Italy: - Villa Monastero; 2003.
[6] GENIA 2003. http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/.
[7] PennTreebank-2 2003. Penn Treebank-2 corpus. www.TB-2.upenn.edu.
[8] Marcus M, Santorini B, Marcinkiewicz MA. Building a large annotated corpus of English: the Penn Treebank. In: Computational linguistics, vol. 19; 1993. p. 297–352.
[9] Mikheev A. Automatic rule induction for unknown-word guessing. In: Computational linguistics, vol. 23, No. 3; ACL 1997. p. 405–23.
[10] Pakhomov S. Modeling filled pauses in medical dictations, In: Student papers section of the proceedings of association for computational linguistics 1999 (ACL'99). p. 619–24.
[11] Pakhomov S, Coden A, Chute C. Creating a test corpus of clinical notes manually tagged for part-of-speech information. In: Proceedings of BioNLP workshop at COLING-2004. p. 62–6.
[12] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP-96), Philadelphia; 1996.
[13] Rindflesch TC, Rajan JV, Hunter L. Extracting molecular binding relations from biomedical text. In: Proceedings of the 6th applied natural language processing conference; 2000. p. 188–95.
[14] Schmid H. Treetagger. http://www.ims.uni-stuttgart.de/~schmid/.
[15] Smith, L, Rindflesch, T, Wilbur WJ. MedPost: a part of speech tagger for biomedical text. In: Bioninformatics journal, vol. 1, No. 1; 2004. p. 1–2.
[16] Tateisi Y, Tsujii J. Part-of-speech annotation of biology research abstracts. In: Proceedings of the 4th international conference on language resource and evaluation 2004 (LREC2004). p. 1267–70.
[17] Weischedel R, Meteer M, Schwartz R, Ramshaw L, Palmucci J. Coping with ambiguity and unknown words through probabilistic models. In: Computational linguistics, vol. 19, No. 2; 1994. p. 359–82.
[18] Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University; 1998.
[19] HL-7. http://www.hl7.org.