RC23195 (W0404-146) April 27, 2004
Computer Science

IBM Research Report

Domain-Specific Language Models and Lexicons for Tagging

Anni R. Coden, Serguei V. Pakhamov*, Rie K. Ando,
Patrick H. Duffy*, Christopher G. Chute*

IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

*Division of Medical Informatics Research
Department of Health Sciences Research
Mayo Clinic
Rochester, MN 55905

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Domain-specific language models and lexicons for tagging

Anni R. Coden[1], Serguei V. Pakhomov[2], Rie K. Ando[1], Patrick H. Duffy[2], Christopher G. Chute[2]

[1]IBM, T.J. Watson Research Center, Hawthorne, NY 10532 anni@us.ibm.com	[2]Division of Medical Informatics Research Department of Health Sciences Research Mayo Clinic Rochester, MN 55905 Pakhomov.Serguei@mayo.edu

Abstract

Accurate and reliable part-of-speech tagging is a pre-requisite 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 tagger accuracy. We show that a large annotated general-English corpus is not sufficient for building a 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 tagger model that gives satisfactory accuracy results on a new domain at a relatively small cost.

1 Introduction

Accurate and reliable part-of-speech (POS) tagging is a pre-requisite for many Natural Language Processing (NLP) tasks such as syntactic parsing, feature extraction for classification, semantic representation, among others that, in turn, form the foundation of NLP-based approaches to information retrieval and data mining. Many high precision statistical POS taggers (Brants 2000), (Brill 1993) are available both in the open source and the 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 (PennTreebank-2 2003) which is distributed by the Linguistic Data Consortium (LDC). While 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, typically one expects to find a large number of new or "unknown" 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 performing adaptation of 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 assignment. Our main methodological research question is to uncover the trade-offs in adapting a general-purpose statistical part-of-speech tagger to a medical English sub-domain. We examine and compare two methods of adaptation – one consists of simply adding a lexicon derived from the target domain, the other involves manual annotation of a number of documents from the target domain and adding the annotations to the general English training data.

In the rest of the paper, we will discuss some related work in section 2. A detailed problem description is presented in section 3. In particular, we will present 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 will show and quantify the relation between the corpus used for model building and the test data. In section 4, we will report on a set of experiments using several combinations of the corpora for cross training and testing. Finally, we will also report on a set of experiments with introducing a domain-specific lexicon and compare the results. We will show that a model based on a small domain-specific corpus in conjunction with a general-purpose English corpus improves the accuracy of a tagger. On the other hand, a domain-specific lexicon used together with a model based on general-purpose English has only a small impact but at the fraction of cost in comparison of developing a domain-specific corpus.

2 Related Work

Part-of-speech 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., Cutting et al. 1992; Kupiec 1992; Weischedel et al. 1993; Brants 2000), maximum entropy models (Ratnaparkhi 1996), and transformation-based learning models (Brill 1994).

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, thus, share distributional characteristics. In addition, the size of the training corpus needs be sufficiently large (typically over one million words) for obtaining 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.

The challenge is to achieve as high accuracy when the training corpus and the test data are part of different types of discourse. It is difficult and expensive to develop a domain-specific training corpus. However, one can safely assume that the unknown word rates increases substantially when the training corpus and test data differ in their type. There are several examples in the literature on how unknown words degrade tagger accuracy.

For example, evaluations of Brandts's HMM-based TnT tagger with smoothing and unknown word prediction modules show an overall accuracy of 96.7% on both NEGRA corpus of German and Penn Treebank of general English corpora (Brants 2000). 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 (Mikheev, 1997). 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% (Brandts, 2000). From these 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.

Rindflesh et al. (2000) report 93.1% accuracy achieved with the Xerox (Cutting et a., 1992) 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 tag-set, which makes comparisons difficult without reducing the Penn Treebank tagset to the SPECIALIST tagset. Smith et al. (2004) 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-specific lexicon in combination with a domain-specific corpus data for training HMM-based taggers such as MedPost happen to be more beneficial that 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.).

Another example of tagger adaptation to the biomedical domain is reported by Jensen et al. (2003). In their work on using biomedical literature for knowledge discovery, Jensen et al. report the results of re-training a TreeTagger (Schmidt, http://www.ims.uni-stuttgart.de/~schmid/) on the GENIA corpus. The tagger trained on 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 on 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 was used for training and testing. However, the results indicate

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

that domain adaptation results in improved performance.

3 Problem Description

The problem we are trying to address is how to adapt a part-of-speech tagger based on general English model to the biomedical domain. We will 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 to the characteristics of the Penn Treebank-2 corpus to gather insight how a model should be built to obtain good accuracy in a part-of-speech tagger. The characteristics we are focusing on are those typically used by taggers.

Our experiments use an HMM tagger since it trains and tests fast and has been shown to be highly accurate. A traditional HMM model for part-of-speech tagging assumes that word emissions are conditioned on tags, and that tags are conditioned on the immediately preceding *n* tags, where n defines the order of the HMM. That is, lexical probabilities p(word|tag) and transition probabilities p(tag| previous *n* tag(s)) are estimated from the training data. In addition, our in-house HMM tagger, used for the experiments reported in this paper, estimates p(word|tag) for unseen and low frequency words from p(ending|tag) for up to 4 characters, p(char-type|tag), and p(unseen|tag), similarly to (Weischedel et al. 1993). The order of the HMM model in the tagger is an input parameter. The simplest model is the uni-gram model without unknown word processing. In that case, the tagger assigns the most frequent tag in the model corpus to unknown words in the test data.

3.1 The Corpora

Our experiments involve three corpora, the Penn Treebank-2 (Marcus, 1993) corpus, the GENIA (2003) corpus and MED, a corpus of clinical notes. In particular, are using a subset (hereafter TB-2) of the Penn Treebank corpus that consists of the Brown and Wall Street Journal collections distributed by the LDC. It is a large, manually annotated with part-of-speech tags corpus, and is widely used to train taggers.

The GENIA corpus (Genia 2003) is a set of 2000 Medline abstracts obtained by using three different search key words. This corpus has also been manually edited for POS tags (Tateisi and Tsujii, 2004), however the guidelines differed slightly from those used for TB-2 and MED (http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/). In particular, proper noun tags are not used in annotating the GENIA collections except for bibliographical information (e.g., author, research institute) and the SYM tag was intentionally used sparsely.

A medical institution developed the proprietary MED corpus. It is the goal of the medical institution to tag their ever-growing set of clinical notes with POS information. The current size of the 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 part-of-speech tags. Three domain experts familiar with the language of the clinical notes annotated the collection. The following is a passage from a typical clinical note:

The patient is a 62 year - old woman diagnosed as having rheumatoid arthritis that was made approximately four years ago. Depression, anemia, hypertension not treated with medications, status post venous stripping, status post hysterectomy and oophorectomy, rheumatoid arthritis.

In contrast, a part of a PubMed article from the GENIA collection is shown here.

TI - IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase.

Although it seems intuitive that these two passages are quite different from each other and from a newspaper article (the Penn Treebank-2), we will qualify and quantify their similarities and differences. Table 1 shows some basic statistics of these corpora. All tokens are made lower case before being counted.

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

Table 1: Size of corpora

The tagger used in our studies, uses number normalization to increase the frequency of low frequency words: each occurrence of a digit is mapped to the digit 0. For example, the number 3 is mapped to 0, 33 is mapped to 00 and L3 is mapped into L0. Table 2 shows the percentage decrease when number normalization is performed. The biggest drop is seen in the GENIA corpus which indicates that many tokens differ only in digits. For example, “#-fold” where # is a one or more digits appears frequently.

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

Table 2: Percentage decrease of number of types due to number normalization

In the rest of the paper, the type counts are for not number-normalized types.

Next, we examined the size of the vocabulary in the three corpora. Towards this end, the number of types in the first 100 000 tokens in each corpus was counted and the results are shown in table 3.

#tokens	TB-2 #types	MED #types	GENIA #types
100,000	11,516	8,691	8,422

Table 3: Size of vocabulary

It is not surprising that the GENIA collection has the smallest vocabulary, as its documents in are the result of a focused search (three keywords). One cannot expect the vocabulary to stay small over all PubMed articles.

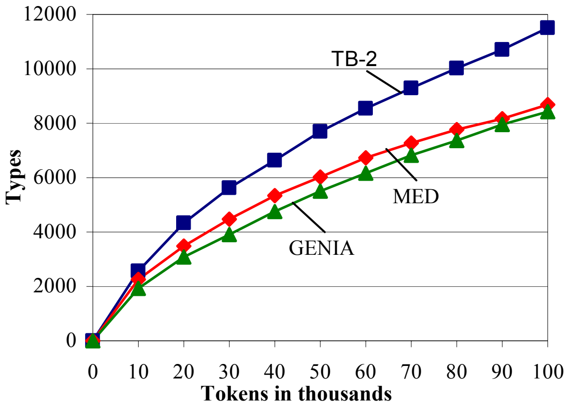

Figure 1:Trend in vocabulary size

Although, Figure 1 shows the trend only up to 100K tokens, one can observe that the gradient of increase in types is smaller for the MED and GENIA collections than for the TB-2 collection. It is not clear whether there is an asymptotic value for the size of vocabularies. The Oxford Dictionary has approximately 600 000 entries, not counting the variants of words and proper names. Hence, looking at the gradient is more meaningful.

Another characteristic to investigate is the sentence length. Table 4 shows the average sentence length in the three corpora. These numbers show that the MED corpus consists of much shorter sentences than the other corpora.

	# token	# sentences	token/sentence ratio
TB-2	128,921	53,362	24.16
MED	100,650	7,299	13.79
GENIA	501,062	18,436	27.18

Table 4: Average sentence length

Our observation of the data in this corpus suggests that a portion of the sentences in the MED corpus consists of sentence fragments that are missing the explicit mention of the subject when for instance the sentence is about the patient. For instance, a note may contain the following sentence fragment: “*Winters in Florida.*” A human can deduce easily from the context that “*Winters*” is a verb, however an automatic POS tagger may have problems correctly tagging this word.

Other corpus characteristics used in POS tagging algorithms are tag distributions and tag transitions. Figure 2 shows the tag distribution by tag groups, tags which start with “N” are grouped together, as are “J”, “V” and “R” tags in the Penn Treebank tag set.

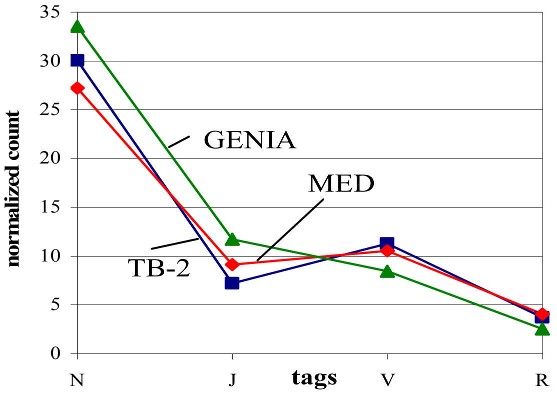

Figure 2: Tag distribution

It is important to take the different guidelines in tagging into account. In particular, the GENIA collection has very few proper noun tags (by design). The proper nouns in the GENIA corpus are tagged in general as nouns. GENIA has a higher percentage of nouns and a lower percentage of verbs. The distributions of tags in TB-2 and MED are quite similar.

Taggers use transition statistics to determine the accurate tag for ambiguous tagged words and for unknown words. Table 5 shows the normalized count of the five most frequent transitions in each corpus.

TB-2	Count	MED	Count	GENIA	Count
DT*NN	100.00	JJ*NN	100.00	NN*NN	100.00
NNP*NNP	90.69	NN*IN	79.82	JJ*NN	99.03
NN*IN	79.30	DT*NN	75.75	NN*IN	88.39
JJ*NN	73.69	NN*.	68.85	IN*NN	57.00
IN*DT	72.72	NN*NN	68.04	DT*NN	53.60

Table 5: Tag transitions

The transition statistics (Table 5) in conjunction with other corpora statistics lead to some more observations. The transition between determiners and nouns is much higher in TB-2 than the other corpora. However, the percentage of tokens 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. Since 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 in the top 5 transitions for that corpus.

3.2 Similarities and Differences of corpora

The goal of this study is to quantify how well a tagger developed for one domain performs on a different domain. In case the accuracy is not satisfactory on a new domain, can it be corrected with a relatively small domain-specific POS tagged corpus or a domain-specific lexicon?

In general, the percentage of out-of-vocabulary words affects the accuracy of a part-of-speech tagger. Table 6 depicts the overlap between the corpora in terms of percentages. It shows the overlap between TB-2 and MED to be approximately 55% of the MED vocabulary. The overlap between TB-2 and GENIA is 63% of the GENIA vocabulary However, only 37% of the GENIA collection overlaps with the TB-2 corpus. Approximately 37% of the MED corpus overlaps with the GENIA collection.

Corpus 1 C1	TB-2	TB-2	MED
Corpus 2 C2	MED	GENIA	GENIA
Overlap % of C2	55.85	63.11	37.09

Table 6: Overlap of types between corpora

Figure 3 shows the number of distinct types in each of the corpora and their mutual overlap. Only 2626 distinct types are present in all three corpora. Adding the GENIA corpus to the TB-2 corpus to build a model for tagging the MED collection should not help much as only 603 new types are added.

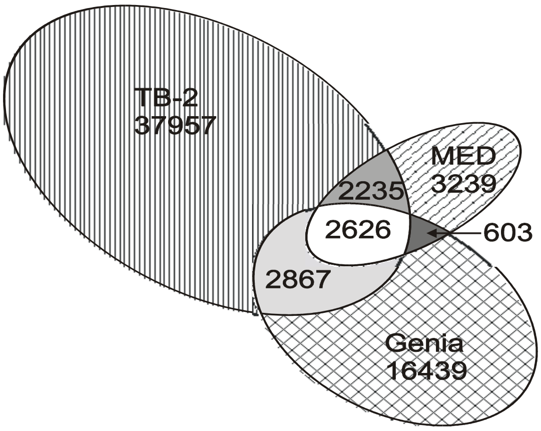

Figure 3: Overlap of types between corpora

We also explore the use of a lexicon to improve the performance of a tagger. The results of this study are discussed in the next section. To gain some insight into the type of lexicon that would most likely be advantageous, we examined the tags in the part of each corpora that does not intersect with any of the other corpora. In particular, more than half of the MED unique vocabulary items are nouns. Contrasting this finding is that the most frequent vocabulary items in the GENIA corpus that are unique to GENIA are nouns (39%) and adjectives (36%).

In general, creating a bigger corpus for training should reduce the out-of-vocabulary rate. However, adding a corpus can also decrease the accuracy if the tag set associated with a word in the additional corpus differs from the tag-set associated with the same word in the original corpus. For example, the word “cold” could be tagged as only an adjective in a general purpose English corpus. In contrast, a medical corpus would tag “cold” both as an adjective and as a noun.

Corpus 1	TB-2	TB-2	MED
Corpus 2	MED	GENIA	GENIA
% types with diff tag-set	51.09	43.77	38.88

Table 7: Ambiguity

The MED corpus adds the biggest percentage of types which have a different tag-set than within the

TB-2 corpus as shown in table 7. Building a model based on the TB-2 and the MED corpus and testing it on the GENIA corpus performs worse than a model just based on the TB-2 corpus.

4 Adaptation Study

To establish a baseline, we trained the tagger with a part of each corpus and tested it on the remaining part.

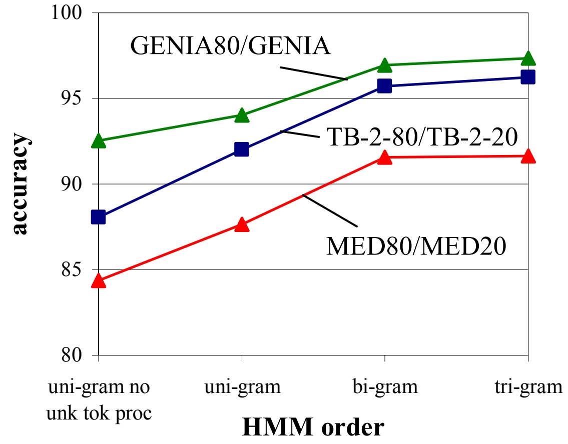

Figure 4: Model build on 80% of corpus, test data is remaining 20% of corpus

In particular, each corpus was split randomly in the 80/20 fashion, where the 80 percent was used for training. This split was done 10 times and the accuracy numbers reported are the average over the 10 runs. All the runs where done with four different underlying models in the tagger. The accuracy did not increase at all, or just very slightly when the order of the Markov Model was set to 4, hence these numbers are omitted from the tables and charts. The most primitive model for the tagger is a uni-gram model without any unknown token processing as described earlier in the paper. (Note: the labels on the graph refer throughout the paper to the model/test data split.)

Table 8 shows the accuracy numbers underlying Figure 4. It shows that changing the order of the model in the tagger improves the accuracy. However, the gains level off with increasing order.

	TB-2	MED	GENIA
uni-gram no unk tok proc	88.08	84.38	92.54
uni-gram	92.01	87.63	94.04
bi-gram	95.69	91.57	96.95
tri-gram	96.24	91.63	97.37

Table 8: Tagger accuracy within a single domain

It is surprising at first to see how well the tagger performs with a base-line model (uni-gram, no unknown token processing) on the GENIA collection. The accuracy numbers can be explained by examining the average out-of-vocabulary rate as shown in table 9.

	TB-2	MED	GENIA
% OOV	3.66	10.18	4.32

Table 9: Out-of –vocabulary rate

Another factor is the high percentage of unambiguous types in the collection. (Unambigous is defined here as having a single tag associated with a word within a single corpus.) Hence a lexicon does nearly as well as a tagger model which takes transitions into account

The MED corpus has a very high out-of-vocabulary rate in comparison to the other corpora. The out-of-vocabulary rate for the MED corpus fluctuates only minimally between each of the 10 runs. This indicates, that although the vocabulary is relatively small, it is evenly distributed throughout the collection.

Most applications will require a higher accuracy than the 92% achieved with a tri-gram model on the MED data. One could project that a higher accuracy could be achieved with a bigger MED corpus, a costly proposition.

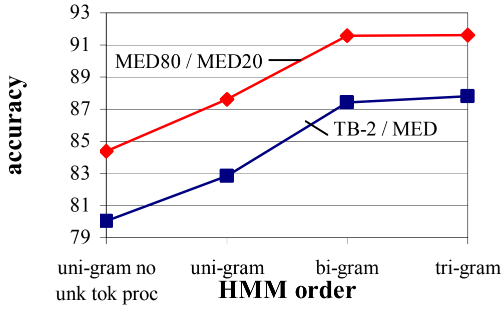

Figure 5: Accuracy on MED test data

Figure 5 shows that the accuracy on the MED data actually decreases when the tagger is trained with TB-2 data only. The out-of-vocabulary rate, which is 10.18% when the tagger is trained with a portion of the MED corpus, increases to 12.47% when the training data is only the TB-2 corpus. Another reason for the decrease in accuracy are the differences in tag distributions and tag-transition distributions.

The differences are even more dramatic when tagging the GENIA collection with a tagger model based solely on TB-2 as shown in Figure 6.

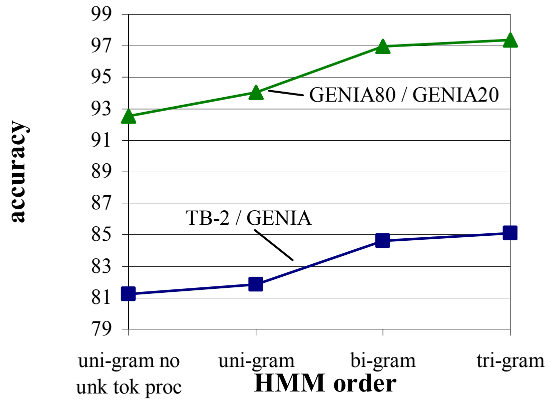

Figure 6: Accuracy on GENIA test data

The out-of-vocabulary rate is 4.32% when the tagger model is based on a portion of the GENIA corpus. The rate increases on average to 21.24% when the tagger model is based on the TB-2 corpus only.

The question arises whether adding the GENIA (or MED) corpus to the TB-2 corpus for training purposes and testing on MED (or GENIA) improves the accuracy.

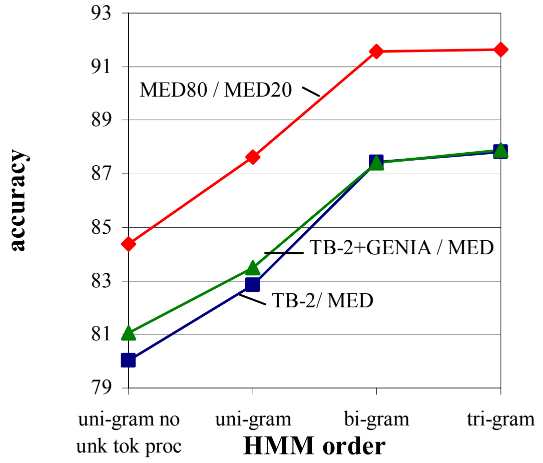

Figure 7: Adding GENIA corpus to TB-2 corpus for model building

Although, adding the GENIA corpus to the TB-2 corpus and testing on the MED corpus improves the performance slightly for some tagger models, the improvement is quite small. This is not surprising, as the GENIA corpus and MED corpus share only a few tokens. The results are nearly identical when adding the MED corpus to the TB-2 corpus for training and testing on the GENIA corpus.

The question arises whether adding a domain-specific corpus to a large general English corpus would improve the accuracy over training with a domain-specific corpus alone. Towards this end, we again randomly split the MED (GENIA) corpus into an 80% training part and 20% testing part. This was done 10 times. The TB-2 corpus was used with the MED corpus to train and then applied on the remaining 20%. The averaged results are shown in Figure 8.

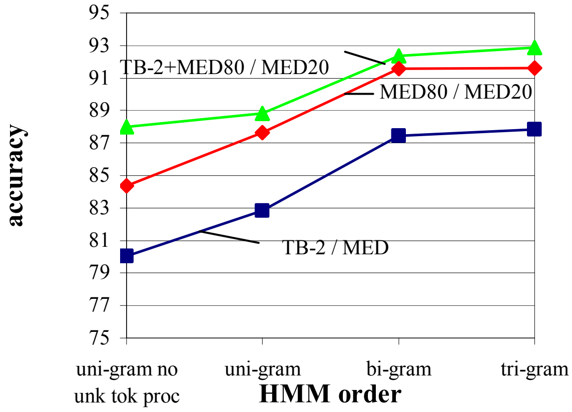

Figure 8: Adaptation with MED domain corpus

Here we see, that adaption with a domain-specific corpus improves the performance. The change is more pronounced for the low-order models in the tagger.

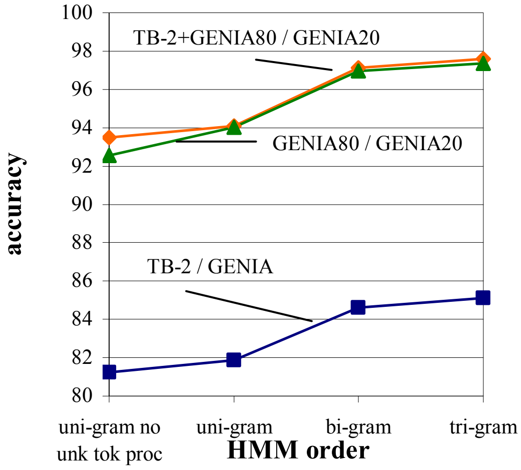

Figure 9: Adaptation with GENIA domain corpus

However, adding the TB-2 corpus to the GENIA corpus changes the accuracy minimally. This is explained by the out-of-vocabulary rate dropping to approximately 2% once a domain-specific corpus is added. One has to note that he GENIA corpus is five times bigger than the MED corpus. What would the accuracy be if the GENIA corpus is restricted to 100K tokens? Can a domain-specific lexicon be used instead? We addressed the second question. We computed the 500 most frequent types from the pool of 16 million clinical notes collection and removed function words. It is noteworthy, that 482 out of the 500 tokens in the lexicon were in the MED corpus. This is indication that the vocabulary in the sampling of clinical notes in the MED corpus is representative of the general collection. We built 5 lexicons (100 … 500) tokens and used each of them in conjunction with a model based on TB-2.

	TB-2	+L100	+L200	+L300	+L400	+L500
uni-gram no unk tok proc	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

Table 10: Adaptation with lexicon

Even a small lexicon improves the accuracy of the tagger over using it without any domain knowledge as shown in table 10. The accuracy improvement grows with the size of the lexicon.

5 Conclusion

Part-of-speech tagging forms a basis for many different natural language applications. Smith (2003) 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 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 based on a general-English corpus in conjunction with a small domain-specific corpus. To achieve the same accuracy on the GENIA corpus, the model has to be built based on (part of) the GENIA corpus, and adding a general-English corpus to built the model does not change the accuracy of the tagger. However, using a domain-specific 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-specific corpus.

References

Thorsten Brants. 2000, TnT – "*A Statistical Part-of-Speech Tagger*", Proceedings of the Sixth Applied Natural Language Processing Conference, ANLP-2000. 224-231

Eric Brill, 1993. "*A Corpus-Based Approach to Language Learning*", Ph.D. Dis, Dep. of Computer and Information Science, University of Pennsylvania

Eric Brill 1994. "*Some Advances in Rule-Based Part of Speech Tagging*", Proceedings of AAAI-94.

Doug Cutting, Julian Kupiec, Jan Pealersen, Penelope Sibun. 1992. "*A Practical Part-of-Speech Tagger*". Proceedings of ANLP-92.

R. Garside, and N. Smith, 1997, "*A hybrid grammatical tagger: CLAWS4*", Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102-121.

Jensen, L., Saric, J., and Bork, P. 2003, "*Utilizing literature for biological discovery*", Proc. of E-BioSci/ORIEL 2003 Villa Monastero, Varenna, Italy.

GENIA 2003, (http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/)

PennTreebank-2 2003. Penn Treebank-2 corpus (**www.TB-2.upenn.edu**)

Marcus, M., B. Santorini, and M. A. Marcinkiewicz 1993, "Building a large annotated corpus of English: the Penn Treebank". Computational Linguistics 19, 297-352.

Andrei Mikheev: 1997, "Automatic Rule Induction for Unknown-Word Guessing" Computational Linguistics 23(3): 405-423

Adwait Ratnaparkhi. 1996. "*A Maximum Entropy Model for Part-of-Speech Tagging*", Proceedings of EMNLP-96.

Rindflesch, T. C., Rajan, J.V., Hunter, L. (2000), "*Extracting Molecular Binding Realtions from Biomedical Text*", Proc. of the 6th Applied Natural Language Processing Conference. 188-195.

Beatrice Santorini: "*Part-of-Speech Tagging Guidelines for the Penn Treebank Project.*", 1991, Technical Report. Department of Computer and Information Science, University of Pennsylvania, Mar. 1991.

Smith, L., Rindflesch, T., Wilbur, W.J. (2003). "MedPost: A Part of Speech Tagger for Biomedical Text". Bioninformatics Journal, Vol 1, no. 1, 1-2.

Ralph Weischedel, Marie Meteer, Richard Schwartz, Lance Ramshaw, Jeff Palmucci. 1993., "*Coping with Ambiguity and Unknown Words through Probabilistic Models*", Computational Linguistics, 19(2):359-382.

Yuka, T. , Tsujii, J. 2004. *"Part-of-Speech Annotation of Biology Research Abstracts."* Proc. LREC 2004.