Creating a Test Corpus of Clinical Notes Manually Tagged for Part-of-Speech Information

Serguei PAKHOMOV
Division of Medical Informatics Research, Mayo Clinic
Rochester, MN
Pakhomov.Serguei@mayo.edu

Anni CODEN
IBM, T.J. Watson Research Center,
Hawthorne, NY 10532
anni@us.ibm.com

Christopher CHUTE
Division of Medical Informatics Research, Mayo Clinic Rochester, MN
Chute@mayo.edu

Abstract

This paper presents a project whose main goal is to construct a corpus of clinical text manually annotated for part-of-speech information. We describe and discuss the process of training three domain experts to perform linguistic annotation. We list some of the challenges as well as encouraging results pertaining to inter-rater agreement and consistency of annotation. We also present preliminary experimental results indicating the necessity for adapting state-of-the-art POS taggers to the sublanguage domain of medical text.

1 Introduction

Having reliable part-of-speech (POS) information is critical to successful implementation of Natural Language Processing (NLP) techniques for processing unrestricted text in the biomedical domain. State-of-the-art automated POS taggers achieve accuracy of 93% - 98% and the most successful implementations are based on statistical approaches to POS tagging. Taggers based on Hidden Markoff Model (HMM) technology currently appear to be in the lead. The prime public domain examples of such implementations include the Trigrams'n'Tags tagger (Brandts 2000), Xerox tagger (Cutting et al. 1992) and LT POS tagger (Mikheev 1997). Maximum Entropy (MaxEnt) based taggers also seem to perform very well (Ratnaparkhi 1996, Jason Baldridge, Tom Morton, and Gann Bierner http://maxent.sourceforge.net).

One of the issues with statistical POS taggers is that most of them need a representative amount of hand-labeled training data either in the form of a comprehensive lexicon and a corpus of untagged data or a large corpus of text annotated for POS or a combination of the two. Currently, most of the POS tagger accuracy reports are based on the experiments involving Penn Treebank data (Marcus, 1993). The texts in Treebank represent the general English domain. It is not entirely clear how representative the general English language vocabulary and structure are of a specialized subdomain such as clinical reports.

A well-recognized problem is that the accuracy of all current POS taggers drops dramatically on unknown words. For example, while the TnT tagger performs at 97% accuracy on known words in the Treebank, the accuracy drops to 89% on unknown words (Brandts, 2000). 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 unknown word "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 Penn Treebank to sublanguage domains with vocabulary that is substantially different from the one represented by the Penn Treebank corpus.

Based on the observable differences between the clinical and the general English discourse and POS tagging accuracy results on unknown vocabulary, it is reasonable to assume that a tagger trained on general English may not perform as well on clinical notes, where the percentage of unknown words will increase. To test this assumption, a "gold standard" corpus of clinical notes needs to be manually annotated for POS information. The issues with the annotation process constitute the primary focus of this paper.

We describe an effort to train three medical coding experts to mark the text of clinical notes for part-of-speech information. The motivation for using medical coders rather than trained linguists is threefold. First of all, due to confidentiality restrictions, in order to develop a corpus of hand labeled data from clinical notes one can only use personnel authorized to access patient information. The only way to avoid it, is to anonymize the notes prior to POS tagging which in itself is a difficult and expensive process (Ruch et al. 2000). Second, medical coding experts are well familiar with clinical discourse, which helps especially with annotating medicine specific vocabulary. Third, the fact that POS tagging can be viewed as a classification task makes the medical coding experts highly suitable because their primary occupation and expertise is in classifying patient records for subsequent retrieval.

We show that, given a good set of guidelines, medical coding experts can be trained in a limited amount of time to perform a linguistic task such as POS annotation at a high level of agreement on both clinical notes and Penn Treebank data. Finally, we report on a set of training experiments performed with the TnT tagger (Brandts, 2000) using the Penn Treebank as well as the newly developed medical corpus..

2 Annotation

Prior to this study, the three annotators who participated in it had a substantial experience in coding clinical diagnoses but virtually no experience in POS markup. The training process consisted of a general and rather superficial introduction to the issues in linguistics as well as some formal training using the POS tagging guidelines developed by Santoriny (1991) for tagging Penn Treebank data. The formal training was followed by informal discussions of the data and difficult cases pertinent to the clinical notes domain which often resulted in slight modifications to the Penn Treebank guidelines.

The annotation process consisted of preprocessing and editing. The pre-processing includes sentence boundary detection, tokenization and priming with part-of-speech tags generated by a MaxEnt tagger (Maxent 1.2.4 package (Baldridge et al.)) trained on Penn Treebank data. Automatically annotated notes were then presented to the domain experts for editing.

3 Annotator agreement

In order to establish reliability of the data, we need to ensure internal as well as external consistency of the annotation. First of all, we need to make sure that the annotators agree amongst themselves (internal consistency) on how they mark up text for part-of-speech information. Second, we need to find out how closely the annotators generating data for this study agree with the annotators of an established project such as Penn Treebank (external consistency). If both tests show relatively high levels of agreement, then we can safely assume that the annotators in this study are able to generate part-of-speech tags for biomedical data that will be consistent with a widely recognized standard and can work independently of each other thus tripling the amount of manually annotated data.

3.1 Methods

Two types of measures of consistency were computed – absolute agreement and Kappa coefficient. The absolute agreement (*Abs Agr*) was calculated by dividing the total number of times all annotators agreed on a tag over the total number of tags.

Kappa coefficient is given in (1) (Carletta 1996)

$$(1) \qquad Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of times the annotators actually agree and P(E) is the proportion of times the annotators are expected to agree due to chance[3].

The Absolute Agreement is most informative when computed over several sets of labels and where one of the sets represents the “authoritative” set. In this case, the ratio of matches among all the sets including the “authoritative” set to the total number of labels shows how close the other sets are to the “authoritative” one. The Kappa statistic is useful in measuring how consistent the annotators are compared to each other as opposed to an authority standard.

3.2 Annotator consistency

In order to test for internal consistency, we analyzed inter-annotator agreement where the three annotators tagged the same small corpus of clinical dictations.

File ID	Abs agr.	Kappa	N Samples
1137689	93.24%	0.9527	755
1165875	94.59%	0.9622	795
1283904	89.79%	0.9302	392
1284881	90.42%	0.9328	397
1307526	84.43%	0.8943	347
Total			**2686**
Average	**90.49%**	**0.9344**	

Table 1. Annotator agreement results based on 5 clinical notes

[3] A very detailed explanation of the terms used in the formula for Kappa computation as well as concrete examples of how it is computed are provided in Poessio and Vieira (1988).

The results were compared and the Kappa-statistic was used to calculate the inter-annotator agreement. The results of this experiment are summarized in Table 1. For the absolute agreement, we computed the ratio of how many times all three annotators agreed on a tag for a given token to the total number of tags.

Based on the small pilot sample of 5 clinical notes (2686 words), the Kappa test showed a very high agreement coefficient – 0.93. An acceptable agreement for most NLP classification tasks lies between 0.7 and 0.8 (Carletta 1996, Poessio and Vieira 1988). Absolute agreement numbers are consistent with high Kappa as they show an average of 90% of all tags in the test documents assigned exactly the same way by all three annotators.

The external consistency with the Penn Treebank annotation was computed using a small random sample of 939 words from the Penn Treebank Corpus annotated for POS information.

Annotator	Abs agr
A1	88.17%
A2	87.85%
A3	87.85%
Average	**87.95%**

Table 2. Absolute agreement results based on 5 clinical notes with an "authority" label set.

The results in Table 2 show that the three annotators are on average 88% consistent with the annotators of the Penn Treebank corpus.

3.3 Descriptive statistics for the corpus of clinical notes

The annotation process resulted in a corpus of 273 clinical notes annotated with POS tags. The corpus contains 100650 tokens from 8702 types distributed across 7299 sentences. Table 3 displays frequency counts for the top most frequent syntactic categories.

Category	Count	% total
NN	18372	18%
IN	8963	9%
JJ	8851	9%
DT	6796	7%
NNP	4794	5%

Table 3 Syntactic category distribution in the corpus of clinical notes.

The distribution of syntactic categories suggests the predominance of nominal categories, which is consistent with the nature of clinical notes reporting on various patient characteristics such as disorders, signs and symptoms.

Another important descriptive characteristic of this corpus is that the average sentence length is 13.79 tokens per sentence, which is relatively short as compared to the Treebank corpus where the average sentence length is 24.16 tokens per sentence. This supports our informal observation of the clinical notes data containing multiple sentence fragments and short diagnostic statements. Shorter sentence length implies greater number of inter-sentential transitions and therefore is likely to present a challenge for a stochastic process.

4 Training a POS tagger on medical data

In order to test some of our assumptions regarding how the differences between general English language and the language of clinical notes may affect POS tagging, we have trained the HMM-based TnT tagger (Brandts, 2000) with default parameters at the tri-gram level both on Penn Treebank and the clinical notes data. We should also note that the tagger relies on a sophisticated "unknown" word guessing algorithm which computes the likelihood of a tag based on the N last letters of the word, which is meant to leverage the word's morphology in a purely statistical manner.

The clinical notes data was split at random 10 times in 80/20 fashion where 80% of the sentences were used for training and 20% were used for testing. This technique is a variation on the classic 10-fold validation and appears to be more suitable for smaller amounts of data.

We conducted two experiments. First, we computed the correctness of the Treebank model on each fold of the clinical notes data. We tested the Treebank model on the 10 folds rather than the whole corpus of clinical notes in order to produce correctness results on exactly the same test data as would be used for validation tests of models build from the clinical notes data. Then, we computed the correctness of each of the 10 models trained on each training fold of the clinical notes data using the corresponding testing fold of the same data for testing.

Split	Hits	Total	Correctness
Average	**21826.3**	**24309**	**89.79%**

Table 4 Correctness results for the Treebank model.

Correctness was computed simply as the percentage of correct tag assignments of the POS tagger (hits) to the total number of tokens in the test set. Table 4 summarizes the results of testing the Treebank model, while Table 5 summarizes the testing results for the models trained on the clinical notes.

The average correctness of the Treebank model tested on clinical notes is ~88%, which is considerably lower than the state-of-the-art performance of the TnT tagger - ~96%. Training the tagger on a relatively small amount of clinical notes data brings the performance much closer to the state-of-the-art – ~95%.

Split	Hits	Total	Correctness
Average	23018.4	24309	94.69%

Table 5 Correctness results for the clinical notes model.

5 Discussion

The results of this pilot project are encouraging. It is clear that with appropriate supervision, people who are well familiar with medical content can be reliably trained to carry out some of the tasks traditionally done by trained linguists.

This study also indicates that an automatic POS tagger trained on data that does not include clinical documents may not perform as well as a tagger trained on data from the same domain. A comparison between the Treebank and the clinical notes data shows that the clinical notes corpus contains 3,239 lexical items that are not found in Treebank. The Treebank corpus contains over 40,000 lexical items that are not found in the corpus of clinical notes. 5,463 lexical items are found in both corpora. In addition to this 37% out-of-vocabulary rate (words in clinical notes but not the Treebank corpus), the picture is further complicated by the differences between the n-gram tag transitions within the two corpora. For example, the likelihood of a DT → NN bigram is 1 in Treebank and 0.75 in the clinical notes corpus. On the other hand, JJ → NN transition in the clinical notes is 1 but in the Treebank corpus it has a likelihood of 0.73. This is just to illustrate the fact that not only the "unknown" out-of-vocabulary items may be responsible for the decreased accuracy of POS taggers trained on general English domain and tested on the clinical notes domain, but the actual n-gram statistics may be a major contributing factor.

6 Conclusion

Several questions remain unresolved. First of all, it is unclear how much domain specific data is enough to achieve state-of-the-art performance on POS tagging. Second, given that it is somewhat easier to develop lexicons for POS tagging than to annotate corpora, we need to find out how important the corpus statistics are as opposed to a domain specific lexicon. In other words, can we achieve state-of-the-art performance in a specialized domain by simply adding the vocabulary from the domain to the POS tagger's lexicon? We intend to address both of these questions with further experimentation.

7 Acknowledgements

Our thanks go to Barbara Abbot, Pauline Funk and Debora Albrecht for their persistent efforts in the difficult task of corpus annotation. This work has been carried out under the NLM Training Grant # T15 LM07041-19.

References

Baldridge, J., Morton, T., and Bierner, G URL: http://maxent.sourceforge.net

Brandts, T (2000) "TnT – A Statistical Part-of-Speech Tagger." In Proc. NAACL/ANLP-2000.

Carletta, J. (1996). Assiessing agreement on classification tasks: The Kappa statistic. Computational Linguistics, 22(2) pp. 249-254.

Cutting, D., Kupiec, J., Pedersen, J, and Sibun, P. A (1992). Practical POS Tagger. In Proc. ANLP'92.

Jurafski D. and Martin J. (2000). Speech and Language Processing. Prentice Hall, NJ.

Manning, C. and Shutze H. (1999). Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

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

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

Poessio, M. and Vieira, R. (1988). "A corpus based investigation of definite description use" Computational Linguistics, pp 186-215.

Ratnaparkhi A. (1996). A maximum entropy part of speech tagger. In Proceedings of the conference on empirical methods in natural language processing, May 1996, University of Pennsylvania

Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. Proc AMIA Symp. 2000; 729-33.

Santorini B. (1991). Part-of-Speech Tagging Guidelines for the Penn Treebank Project. Technical Report. Department of Computer and Information Science, University of Pennsylvania.

UMLS. (2001). UMLS Knowledge Sources (12th ed.). Bethesda (MD): National Library of Medicine.