Methods Paper ■

Heuristic Sample Selection to Minimize Reference Standard Training Set for a Part-Of-Speech Tagger

Kaihong Liu, MD, MS, Wendy Chapman, PhD, Rebecca Hwa, PhD, Rebecca S. Crowley, MD, MS

Abstract Part-of-speech tagging represents an important first step for most medical natural language processing (NLP) systems. The majority of current statistically-based POS taggers are trained using a general English corpus. Consequently, these systems perform poorly on medical text. Annotated medical corpora are difficult to develop because of the time and labor required. We investigated a heuristic-based sample selection method to minimize annotated corpus size for retraining a Maximum Entropy (ME) POS tagger. We developed a manually annotated domain specific corpus (DSC) of surgical pathology reports and a domain specific lexicon (DL). We sampled the DSC using two heuristics to produce smaller training sets and compared the retrained performance against (1) the original ME modeled tagger trained on general English, (2) the ME tagger retrained on the DL, and (3) the MedPost tagger trained on MEDLINE abstracts. Results showed that the ME tagger retrained with a DSC was superior to the tagger retrained with the DL, and also superior to MedPost. Heuristic methods for sample selection produced performance equivalent to use of the entire training set, but with many fewer sentences. Learning curve analysis showed that sample selection would enable an 84% decrease in the size of the training set without a decrement in performance. We conclude that heuristic sample selection can be used to markedly reduce human annotation requirements for training of medical NLP systems.

■ **J Am Med Inform Assoc.** 2007;14:641–650. DOI 10.1197/jamia.M2392.

Introduction

Natural Language Processing (NLP) applications are used in medical informatics for structuring free-text, for example, for coding information and extracting meaning from medical and scientific documents. Many current, state-of-the-art systems employ machine learning or statistically based approaches that are developed and tested with the general English domain. These systems use models that are corpus-based and trained on large, manually annotated corpora such as Penn Treebank.[1] Accuracy of such NLP components is highly dependent on the degree of similarity between the training set and the documents that will ultimately be processed.

Large corpora of manually annotated medical documents do not currently exist for training Medical NLP applications. Privacy concerns are one barrier to development of corpora. De-Identification systems that automatically remove the HIPAA identifiers[2] can help minimize this barrier. Another significant barrier is that corpora require substantial time and effort from experts to manually annotate documents. Therefore, research in this field has focused on identifying other methods for obtaining training data such as development of domain lexicons—linguistic knowledge bases that cover specific medical domains. In this study, we evaluated heuristic sample selection as a potential method for minimizing the training set requirements for retraining a corpus-based medical NLP component.

Affiliations of the authors: Department of Biomedical Informatics (KL, WC, RSC), Department of Computer Science (RH), Department of Pathology (RSC), University of Pittsburgh School of Medicine, Pittsburgh, PA

The authors acknowledge the contributions of Aditya Nemlekar, and Kevin Mitchell for development of the annotation software, Heather Piwowar, Heather Johnson and Jeannie Yuhaniak for annotation of the SPR corpus, Kaihong Liu was supported by the Pittsburgh Biomedical Informatics Training Grant T15 LM0070759 during the completion of this work. The work was funded by the Shared Pathology Informatics Network U01-CA091343.

Correspondence and reprints: Rebecca Crowley, MD, MS, Department of Biomedical Informatics, University of Pittsburgh School of Medicine, UPMC Shadyside Cancer Pavilion—Room 307, 5230 Centre Avenue, Pittsburgh, PA 15232; e-mail: <crowleyrs@upmc.edu>.

Received for review: 01/31/07; accepted for publication: 05/21/07

Background

Differences between Medical Language and General English

A foundational assumption of statistical NLP taggers is that the probability distribution of words and features used to establish the statistical model remains the same between training data and testing data. The use of these systems with existing models for medical documents is therefore limited by the significant differences of medical language when compared with general English. These differences have been well studied and include:

1. Medical language often contains ungrammatical writing styles. Shorthand and abbreviations are very common.[3,4,5]
2. Institutional variations and individual variations in linguistic construction and formatting are frequent.[5]
3. Distinct sublanguages exist within medicine. For example, different types of reports can show marked structural difference.[6]

4. Medical language often contains a plethora of negations,[7] nouns and prepositional phrases.[8]
5. The size of the medical vocabulary is very large. There are many complex medical terms, organ or disease names, and staging codes.[3,4,9]
6. There is an assumed common body of knowledge between the writer and reader. Therefore details are often left out because the meaning is implicitly understood between experts.[4,10,11]

Part-of-Speech Tagging

Part-of-speech (POS) tagging is an important component for many NLP tasks such as syntactic parsing, feature extraction and knowledge representation. Therefore, POS tagging is the foundation of NLP-based applications. Currently there are several state-of-the-art POS taggers that use machine learning algorithms, including Markov Models,[12,13] probability decision trees,[14,15] and cyclic dependency networks.[16] Other taggers such as the transformation-based tagger or the Brill tagger are primarily symbolic rule-learners and automatically determine the rules from previously tagged training corpora.[17] Ratnaparkhi's[18] Maximum Entropy tagger combines the advantages of all of these methods and has achieved 96.6% accuracy on the Wall Street Journal (WSJ) corpus. All of these taggers have been trained on the WSJ corpus from the Penn Treebank project,[19] and all reported comparable accuracy on WSJ.

There have been previous attempts to develop medical language specific POS taggers. Smith et al. developed MedPost,[12] a POS tagger for biomedical abstract text. They developed a corpus of 5700 manually annotated sentences derived from MEDLINE. MedPost, adapted from a Hidden Markov Model (HMM) tagger, achieved 97.43% accuracy using its native tag set and 96.9% accuracy using the Penn Treebank tag set. However, the high accuracy of MedPost may be specific to the particular medical and scientific sublanguage for which it was developed. Divita et al developed dTagger using the same training set developed by Smith.[20] dTagger incorporates POS information from the SPECIALIST lexicon to identify the POS tag on both single word and multi-word items.

The accuracy of statistical POS taggers trained in general English decreases dramatically when applied to medical language. This is largely due to the high percentage of words that have not been seen by the tagger so that the statistical features used by the tagger to predict POS will be unknown for those words. Smith has observed that a 4% error rate on POS tagging corresponds to approximately one error per sentence.[12] For subsequent components, such as parsers, this error rate may exceed acceptable limits. In order to achieve high accuracy for a statistical tagger, domain specific approaches are required.

Alternatives to Development of Large Domain Specific Corpora

Development of domain specific statistical NLP taggers is limited by the requirement for an annotated corpus. Alternative approaches are needed to minimize the "annotation bottleneck" involved in retraining statistical systems. Coden et al. have studied domain lexicons as an alternative approach.[3] They compared the tagging accuracies of a HMM tagger on three document sets, two of which were medically related (GENIA and MED). GENIA is a set of 2000 MEDLINE abstracts obtained by using three search key words: "Human," "Blood Cells," and "Transcription Factors." MED contains clinical notes dictated by physicians and subsequently transcribed and filed as part of the patients' electronic medical record. As a baseline, they found that the HMM tagger trained on the Penn Treebank performed poorly when applied to GENIA and MED, decreasing from 97% (on general English corpus) to 87.5% (on MED corpus) and 85% (on GENIA corpus).

Coden et al. then compared two methods of retraining the HMM—a domain specific corpus, vs. a 500-word domain specific lexicon. The corpus increased accuracy of the tagger by 6% to 10% over tagging with general English training only. The lexicon increased accuracy of the tagger by 2% over tagging with general English training only. Although the authors noted that the domain-specific lexicon had the advantage of being much less expensive to develop, it appears that use of a domain corpus was superior to a domain lexicon.

Finally, Coden and colleagues studied the effect of training and testing in different domains. They used existing publicly available medical related corpora (e.g., GENIA) in conjunction with Penn Treebank corpus to train the tagger, then used this tagger to tag a set of documents (e.g., MED) which had a slightly different sublanguage than GENIA, although they were all medically related. Using the general English corpus plus the MED corpus in training did improve the tagging accuracy on GENIA, but the general English corpus plus GENIA added only minimal improvement when tested with MED data. They conclude that a training corpus from the same domain as the testing corpus is necessary.

Other recent work also supports the importance of a domain corpus for retraining a POS tagger. Tsuruoka et al.[21] and Tateisi et al.[22] retrained a POS tagger that uses the cyclic dependency network method. In both studies, the tagger was retrained with domain specific corpora derived from MEDLINE, and showed a significant increase in POS tagging accuracy over WSJ alone.

If domain corpora remain a necessity for training statistical taggers, the best solution to the "annotation bottleneck" problem may be to minimize the amount of human annotation required. The idea behind sample selection is to actively learn from a document which information is more helpful to a tagger to build a statistical model for POS tagging.[23,24] Documents that contain these characteristics should be preferentially added to the training set, because this type of document is more informative than others. Documents selected for maximum effect rather than random selection can reduce the labor and expense of manual annotation yet still provide the benefits associated with larger corpora. Hwa has defined the training utility value (TUV) associated with each datum and used this to identify documents that will be included for manual annotation.[24] She has applied sample selection to two syntactic learning tasks: training a prepositional phrase attachment (PP-attachment) model[24] and training a statistical parsing model.[24] In both cases, sample selection significantly reduced the size requirement of the training corpus.

We sought to build on this method by testing a sample selection method based on general heuristics and utilizing publicly available medical language resources. We used a Maximum Entropy (ME) Modeled statistical tagger—a highly accurate Part-Of-Speech (POS) tagger originally trained on the Wall Street Journal corpus. Our document set consisted of surgical pathology reports (SPRs)—clinical documents that describe pathologic findings from biopsies and resections of human tissue. In the future, the heuristics developed in this study could be extended to other NLP components and other medical document types.

Research Questions

We examined six research questions:

1. How do frequencies of parts of speech for pathology reports differ from other medical and general English training sets used for statistical part-of-speech taggers?
2. What is the performance of the MedPost and ME POS taggers on a corpus of pathology reports, without modification of the native training sets?
3. What is the effect on performance of retraining the ME POS tagger with a domain lexicon or a domain specific corpus?
4. Does heuristic sample selection decrease the number of annotated examples needed for retraining and by how much?
5. What is the effect of training set size on performance, for heuristic sample selection?
6. How does retraining affect POS tagging error distribution?

Materials and Methods

Materials

Maximum Entropy Modeled POS Tagger (ME)

We used a publicly available ME tagger[25] for the purposes of evaluating our heuristic sample selection methods. The ME tagger was trained on a large general English corpus—Wall Street Journal articles from the Penn Treebank 3 project that had been previously manually annotated with POS information. The system learns either probability distributions or rules from the training data and automatically assigns POS tags to unseen text. For a given sentence or word sequence, the ME tagger uses features to the model such as prefixes and suffixes of length = 5, as well as whether the word contains a number, hyphen, or an upper-case letter. Therefore, the features that will be considered include the current word, previous two words, two suffix words and two previous words tags. These features are only considered when the feature count is greater than ten. Features occurring less than ten times are classified as rare. Those features occur sparsely in the training set, and it is difficult to predict the behavior of the feature because the statistic may not be reliable. In this case, the model will use heuristics or additional specialized, such as word-specific features.

MedPost Tagger

We used the MedPost tagger[12] as a baseline method for this study. MedPost was trained on 5700 manually annotated sentences randomly selected from MEDLINE abstracts. MedPost is trained on medical language, but is not easily retrainable on specific sublanguages. We reasoned that any adaptation to the ME tagger must at least exceed what could be achieved by MedPost in order to be worth the effort of manual annotation. MedPost tagger can be run with either the SPECIALIST Lexicon tag or Penn Treebank tag sets.

SPECIALIST Lexicon

SPECIALIST is one of the UMLS Knowledge Sources.[26] It provides lexical information for biomedical terms and for general English terms. We used the SPECIALIST lexicon as a source of medical and scientific parts of speech for identifying documents where there are a high frequency of terms that are either (a) unlikely to have been previously encountered by the ME tagger, or (b) ambiguous terms in which the general English part-of-speech may be different from the medical usage.

Surgical Pathology Reports (SPRs)

For training and testing, we drew from a set of 650,000 de-identified surgical pathology reports obtained from the last 10 years of records at the University of Pittsburgh Medical Center. The document set includes cases from multiple University affiliated hospitals. Use of the de-identified SPRs was reviewed by the Institutional Review Board and determined to be exempt (IRB Exemption # 0506031). SPRs were chosen because they represent important medical information that can be used for both clinical and basic sciences.

Methods

To address the research questions, we compiled several data sets for training and testing different POS taggers on the SPR's. All data sets were first tokenized into individual words. We compared reference standard annotations of the reports against automated annotations to compare the performance of various methods. POS tagging (both manual and automated) was performed on individual words.

Data Sets

We generated six data sets in this project to address our research questions (Figure 1). For Data Set 1, we randomly selected 250 SPRs from 650,000 de-identified SPRs available at the University of Pittsburgh Medical Center. Data Set 1 was manually annotated with POS information by trained POS annotators and constituted our *Reference Standard (RS)*. The development of this reference standard is described later. Data Set 1 (which contained 161 unique reports) was split into two parts comprising Data Set 2 and Data Set 3.

Data Set 2 consisted of 20% of the 161 manually annotated SPRs (32 SPRs) from the Reference Standard for use as our *Test Corpus (TC)*. TC was used to measure the performance of POS tagging.

Data Set 3 consisted of the remaining 80% of the annotated RS (129 SPRs) and constituted our *Domain Specific Corpus (DSC)*. This data set was used for development of all DSC-based adaptations. A total of 16,638 words were present in the DSC. The entire DSC was used for retraining the ME tagger in order to obtain the upper bound of accuracy attainable with the entire corpus. Partitions of this data set were used to test the heuristic selection methods and to generate learning curves.

We then developed two heuristics for selection of individual sentences from DSC documents. Heuristics for sample selection are described later. Each heuristic was used to partition the DSC. Data Set 5 reflects data selected using Heuristic 1

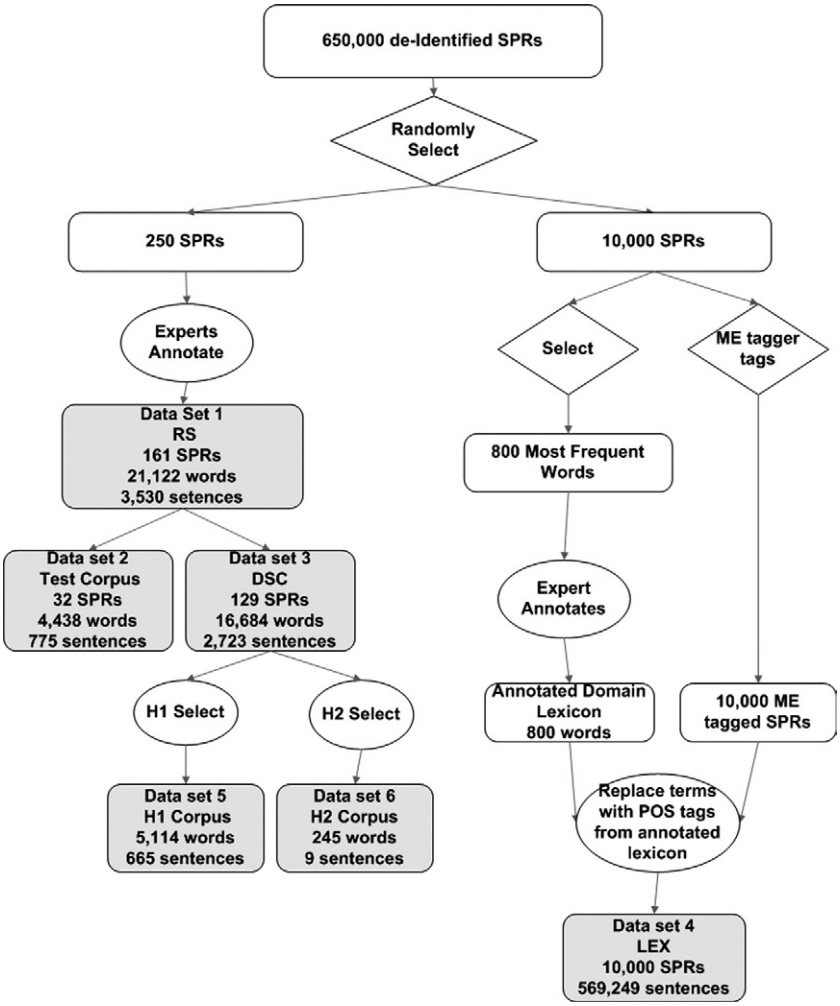

Figure 1. Development of corpus and data sets.

(*H1*) and Data Set 6 reflects data selected using Heuristic 2 (*H2*).

In addition, we randomly selected 10,000 SPRs from the total pool of 650,000 SPRs. From this set of 10,000 reports, we generated a frequency distribution, excluding stop words and determiners. We selected the top 800 words as our domain lexicon which is comparable to the number used by Coden.[3] Each entry could thus represent one or more usages or contexts in the corpus of reports. One pathologist manually annotated each entry with a single POS tag considered to represent the most frequent usage, based on her expertise. This created a simple baseline, analogous to the method used by Coden,[3] that could be compared with more sophisticated methods. To generate Data Set 4 (*LEX*), we first tagged the corpus of 10,000 SPRs with the original ME tagger, then, replaced words found in the 10,000 ME tagger tagged SPRs with POS tags from the domain lexicon.

Development of the Reference Standard (Data Set 1)

In order to achieve the most the reliable reference standard possible, we used an iterative, three-step process where we trained annotators on increasingly difficult tasks and then provided feedback on performance and consistency. First, we trained five annotators and selected the three top-performing annotators to annotate the reference standard. In step two, annotators were given 250 SPRs to annotate. In step three, we collected the manually annotated documents from each annotator and merged them into a single reference standard. We assessed the reliability of the reference standard by calculating absolute agreement between the annotators and agreement adjusted for chance, using the kappa coefficient.

Training reference standard annotators

Five prospective annotators, all with some knowledge of medical language processing, were recruited. Training began with a 30-minute didactic session during which an introduction to the general guidelines was given and all tags were reviewed. This was followed immediately by a one-hour training session, where annotators inspected real examples from the Penn Treebank corpus. Throughout the training of the annotators, the general guidelines for POS tagging developed by Santorini[27] for tagging Penn Treebank data were used. The Penn Treebank POS tag set consists of 36 POS tags.

After the first meeting, there were two rounds of independent, hands-on annotation training. For each round, one standard WSJ document was given to each annotator to test their ability to perform POS tagging. Upon return of the annotated files, we calculated the number of absolute agreements with Penn Treebank annotations. At a follow up meeting, we discussed the problems encountered after each round.

Table 1 ■ Descriptive Statistics of Human Annotated Corpora Used in Study

	Wall Street Journal	MEDLINE Abstracts	Surgical Pathology Reports
Words	1,019,828	155,980	57,565
Word Types	37,408	14,785	3,339
Word Types per 100,000 words	3,668	9,478	5,800
Sentences	48,936	5,700	3,021
Average words per sentence	20.8	27.4	19.05
Average verb per sentence	3.1	2.7	1.05

The three top annotators were selected based on their POS tagging performance by comparing their POS assignments with those in the Penn Treebank after the first training session. In the second training session, the three selected annotators spent two hours annotating a single pathology report which had not been encountered in the initial training. This was followed by a discussion of examples of POS ambiguity. During this session, we reviewed terms with POS that caused difficulty. In many cases, POS ambiguity was resolved by referring to the context in discussion. During the discussion, all disagreements were discussed and corrected. All annotation during Step 1 was performed using a modified Excel spreadsheet.

Generation of the reference standard

Each annotator was then given a set of 250 SPRs to annotate on their own time. During this stage, annotators utilized our modification of the GATE software for annotation. Only the final diagnosis, gross description and comment sections were annotated with POS information since these sections contain almost all of the medical content. The completed annotation sets were then merged into a single "gold standard" using a majority vote if there was disagreement between two of the annotators. When all three annotators disagreed, we randomly selected one of the annotations as the reference standard. During examination of reference standard data, we determined that 89 of the documents were duplications that we had not caught before the annotation. We excluded the duplicates from the reference standard, yielding a final corpus of 161 surgical pathology reports. To measure inter-rater reliability, we calculated the inter-rater agreements and pair-wise Kappa coefficient as described by Carletta.[28]

Comparative Statistics of Human Annotated Corpora

In addition to the Reference Standard we developed, we also had two other manually annotated corpora that were used for comparative purposes—the Wall Street Journal corpus used to train ME tagger, and the MEDLINE corpus used to train MedPost. We examined descriptive statistics comparing these three corpora to determine the distribution and frequencies of POS, in order to:

a) Determine the frequency and distribution of POS in SPRs as compared to general English corpus (WSJ), and MEDLINE abstracts.

b) Develop a list of terms from pathology reports that had not been seen in WSJ.

c) Develop a list of terms that had not been seen in WSJ and were also absent from the Specialist Lexicon.

Heuristics for Sample Selection

We examined the information obtained from the descriptive statistic study and then compiled a list of terms that could help develop heuristic rules for sample selection:

Heuristic 1

We retrained the ME tagger using only selected sentences from the DSC (Data Set 3) where there was a term with high frequency in surgical pathology reports that did not exist in the Wall Street Journal corpus. Terms in this category would be more likely to be tagged in error, because the WSJ had not seen them before. The sentence was used as the base unit for sample selection because the same term can have a different POS depending on its surrounding words and features. The tagger uses this contextual information for disambiguation.

Heuristic 2

We retrained the ME tagger using only selected sentences from the DSC (Data Set 3) where there was a term with high frequency in surgical pathology reports that did not exist in the Wall Street Journal corpus and the Specialist Lexicon. These terms may represent highly specialized medical terminology not covered in the general medical terminology of SPECIALIST.

Evaluation Study

We created four adaptations to the ME tagger by supplementing the existing WSJ training corpus with one of our data sets and retraining the ME tagger, as follows:

1. ME trained with DSC (Data Set 3) + WSJ corpus
2. ME trained with LEX (Data Set 4) + WSJ corpus
3. ME trained with H1 (Data Set 5) + WSJ corpus
4. ME trained with H2 (Data Set 6) + WSJ corpus

We evaluated the four retrained ME taggers described above and compared POS tagging accuracies against two baseline accuracies—(1) the POS tagging accuracy on pathology reports by the ME tagger that had been trained on Penn Treebank data (PT) only, and (2) the POS tagging accuracy on pathology reports by the MedPost tagger that had been trained on MEDLINE abstracts. All evaluation studies were done on Data Set 2 (Test Corpus), which contained 32 SPRs. Single train and test partitions are not reliable estimators of the true error rate. Therefore, during evaluation of sample selection heuristics, we used 10-fold cross validation. We selected training data based on the heuristics from the randomly selected 80% of RS. The remaining 20% of RS was used as the test data set. We report the range of the performance over 10 runs for each training cycle.

Learning Curve Study

We also performed a learning curve study on H1, determining the effect of sample size on accuracy, in order to identify the minimum quantity of training data required to achieve reasonable POS tagging accuracy. The methodology is similar to that described by Hwa[24] and Tateisi.[22] H1 training data were randomly divided into 10 parts. We trained the tagger with a 10% incremental increase of training data over the 10 parts. We did 10 runs for each of 10% training data increment. The 10% training data were randomly selected from total H1 selected sentences for 10 times. The accuracy of POS tagging was measured after each cycle of training.

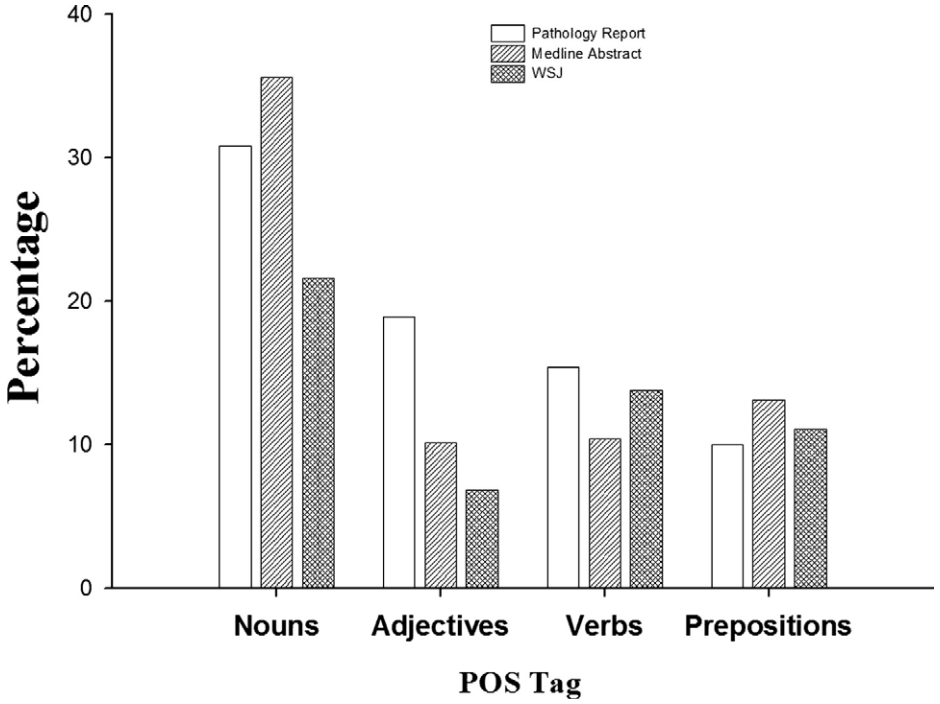

Figure 2. POS distribution comparison for SPR, MEDLINE Abstract, Wall Street Journal corpora.

We report the range of the performance for each incremented training.

POS Tagging Error Analysis

We analyzed POS tagging errors produced by (1) the original ME tagger trained with WSJ alone, (2) the ME tagger retrained with WSJ supplemented by the domain specific corpus (Data Set 3), and (3) the ME tagger retrained with WSJ supplemented by H1 selected domain specific corpus (Data Set 5). For each of these three annotation sets, we determined the distribution of errors compared to the reference standard.

Results

Inter-rater Reliability of Reference Standard

A total of 161 pathology reports were manually annotated by three annotators. The average total annotation time (excluding training time) was 62 hours. The absolute agreement between at least two annotators was 96% and the absolute agreement between three annotators was 68%. The average pair-wise Kappa coefficient was 0.84.

Descriptive Statistics of DSC, WSJ Training Corpus, MEDLINE Training Corpus, and Domain Lexicon

Descriptive statistics regarding the three corpora are shown in Table 1. The percentage of words in the SPR that were not seen in the WSJ was 30%. The relative distribution of nouns, adjectives, verbs and prepositions is shown in Figure 2. Both pathology reports and MEDLINE abstracts contain a higher percentage of nouns when compared to the Wall Street Journal. Pathology reports contain a higher percentage of adjectives and verbs when compared to MEDLINE abstracts. We also compiled the distribution of POS tags for the domain lexicon, which are shown in Figure 3.

Baseline Accuracies of POS Tagging by the Two Taggers

Two baseline accuracies of POS tagging were obtained for purposes of comparison. The accuracy of POS tagging of

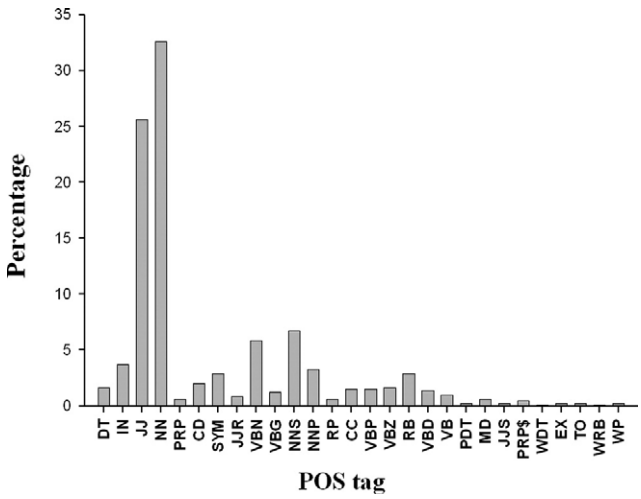

Figure 3. POS distribution for Domain Lexicon. Parts of speech are abbreviated as follows: DT—determiner; IN—preposition or conjunction, subordinating; JJ—adjective or numeral, ordinal; NN—noun, common, singular or mass; PRP—pronoun, personal; CD—numerical, cardinal; SYM—symbol; JJR—adjective, comparative; VBN—verb, past participle; VBG—verb, present participle or gerund; NNS—noun, proper, plural; NNP—noun, proper, singular; RP—particle; CC—numeral, cardinal; VBP—verb, present tense, not 3rd person singular; VBZ—verb, present tense, 3rd person singular; RB—adverb; VBD—verb, past tense; VB—verb, base form; PDT—pre-determiner; MD—modal auxiliary; JJS—adjective, superlative; PRP$—pronoun, possessive; WDT—WH-determiner; EX—existential there; TO—"to" as preposition or infinitive marker; WRB—Wh-adverb; WP—WH-pronoun.

Table 2 ■ Evaluation Results

	Tagger	Accuracy
Baselines	ME	79%
	MedPost	84.20%
Adapted POS Taggers	ME + DSC (3,530 sentences)	93.90%
	ME + LEX	84%
	ME + H1 (665 sentences)	92.70%
	ME+ H2 (9 sentences)	81.20%

SPRs by the ME tagger trained on general English was 79%. The accuracy of POS tagging of SPRs by the Medpost tagger was 84%. In addition to these baselines, we established an upper bound for retraining using sample selection by determining the performance of the ME tagger retrained with the entire DSC. In conjunction with the general English corpus to train ME tagger, use of the entire DSC achieved a 93.9% accuracy of POS tagging.

Accuracies of POS Tagging After the Three Adaptations

Adding a small domain lexicon improved the accuracy of tagging from 79% to 84.2%, which was comparable to the accuracy of the MedPost tagger. Heuristic H1 achieved a substantial increase, from 79% to 92.7% accuracy, nearly matching the upper bound established using the entire DSC. The range of accuracy of H1 over the 10 fold validation was 92.7% ± 0.44. Heuristic H2 produced a smaller improvement over baseline (from 79% to 81%). Table 2 provides a summary of all the evaluation results.

Learning Curve Study

The learning curve (Figure 4) demonstrates improvement in performance from 10 to 50% and a leveling off of performance gains at approximately 50% of the H1 training set. Thus, only half of the words in the H1 training set (total 2557 words) were sufficient to achieve a performance gain nearly equivalent to the entire DSC (16,680 words). This corresponds to an 84% decrease in the size of the corpus that must be annotated with no appreciable decrement on the resulting performance of the POS tagger.

POS Tagging Error Distribution

The most frequent errors in POS assignment are shown in Table 3 (for ME tagger trained on WSJ alone), Table 4 (for ME tagger trained on WSJ and entire domain corpus) and Table 5 (for ME tagger trained on WSJ and H1 selected domain corpus). Each table depicts a 10-by-10 confusion matrix showing the most frequent errors in POS assignment (>90% for each training set). In Table 6, we compare the distribution of POS tagging errors between these three annotation sets. Error analysis shows a decrement in errors across the spectrum of ambiguities when the ME tagger is trained using the domain specific corpus. The distribution of errors for the H1 selected corpus is very similar to the distribution obtained using the entire domain specific corpus.

Discussion

Statistical taggers rely on large, manually annotated data sets as training corpora. The training data set needs to be large in order to learn reliable statistics. The underlying assumption of these statistical taggers is that the probability distribution remains the same between training data and testing data. POS taggers are an example of a statistical tagger frequently used in medical NLP. Common approaches to POS tagging include Hidden Markov Models (HMM) as well as Maximum Entropy models. Rare unknown words are particularly problematic for corpus-based statistical taggers—and it is exactly these rare unknown words which are so frequent in medical documents. For this reason, larger and more diversified data sets are usually necessary to achieve high accuracy.

Each tagging algorithm utilizes some method to deal with unknown words. Some algorithms assume each unknown word is ambiguous among all possible tags, and therefore assigns equal probability to each tag. Some algorithms assume the probability distribution of tags over the unknown word is very similar to the distribution of other

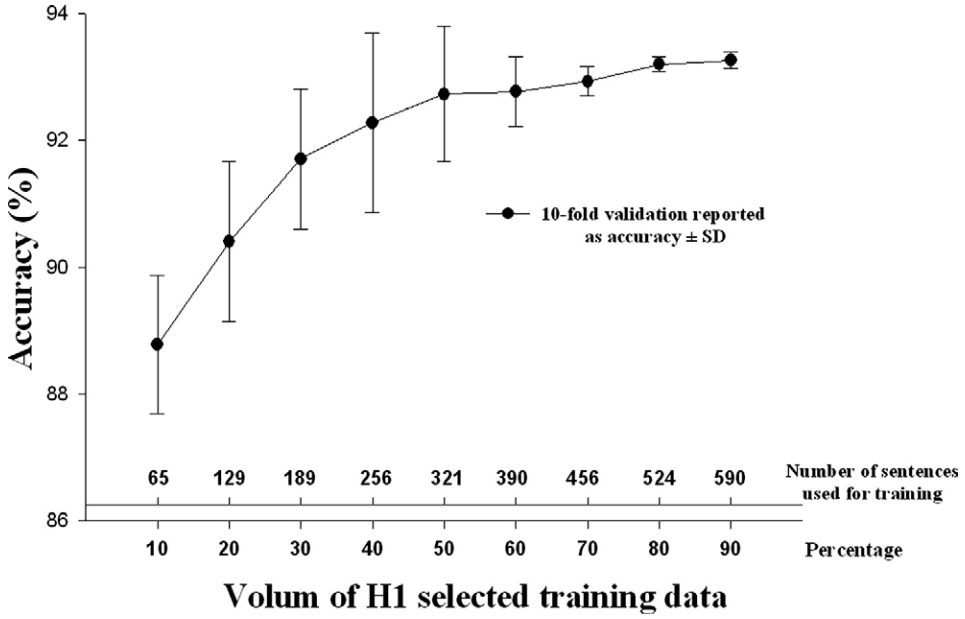

Figure 4. Number of annotated sentences needed to adequately retrain POS tagger.

Table 3 ■ Partial Confusion Matrix Showing Distribution of Most Frequent POS Tagging Errors by ME Tagger Trained with WSJ only

Tagging Error	Reference Standard	ME tagger trained with WSJ only										
		NNP	NN	VBN	CD	NNS	JJ	VBD	VBZ	RB	IN	% of Total
JJ		20% (271)	6% (77)	2%(32)	—	—	—	1% (9)	—	1% (7)	0% (2)	30% (407)
NN		23% (310)	0% (0)	0% (1)	0% (2)	3% (36)	2% (23)	0% (1)	0% (6)	0% (1)	0% (1)	29% (393)
LS		5% (70)	1% (17)	0% (1)	4% (50)	0% (1)	0% (5)	0% (1)	—	—	—	11% (148)
NNS		3% (42)	7% (89)	—	—	—	—	—	1% (7)	—	—	10% (141)
VBD		—	—	4% (57)	—	—	—	—	—	—	—	4% (57)
VBN		1% (12)	0% (1)	—	—	—	0% (6)	1% (10)	—	0% (1)	0% (3)	2% (33)
IN		1% (10)	0% (2)	—	0% (1)	0% (1)	0% (1)	—	—	0% (3)	—	2% (29)
VBZ		0% (1)	0% (1)	0% (3)	0% (1)	1% (14)	0% (1)	—	—	—	0% (2)	2% (26)
CD		2% (21)	0% (1)	—	—	0% (1)	—	—	—	—	—	2% (24)
RB		1% (1)	—	—	—	—	1% (7)	—	—	—	0% (1)	1% (18)
% of Total		57% (768)	15% (202)	7% (95)	4% (55)	4% (53)	4% (52)	2% (21)	1% (19)	1% (16)	1% (11)	95% (1276)

JJ = adjective; NN = noun, singular or mass; LS = list item marker; NNS = noun, plural; VBD = verb, past tense; VBN = verb, past participle; IN = preposition or conjunction, subordinating; VBZ = verb, present tense, 3rd person singular; CD = numeral, cardinal; RB = adverb; NNP = noun, proper, singular Number of errors shown in percentage of total errors, with counts in parentheses, for most frequent errors.

words that occur only once in the training set. They assign average probabilities of words appearing once to all unknown words. More sophisticated algorithms use morphological and orthographic information. For example, words starting with capital letters are likely to be proper nouns. Even with more sophisticated methods, the accuracy of POS tagging an unknown is, at best, 85%.[13,18] These algorithms perform best when the percentage of unknown words is low in the testing data. However, if the tagging data comes from a different domain, the number of unknown words is likely to be quite large. This study showed that more than 30% of words in the SPRs were unknown to the general English trained tagger. It is therefore not surprising that the ME tagger achieved only 79% accuracy.

Statistical taggers make use of contextual features surrounding the word to be tagged, so differences in corpora syntactic structure provides a second reason why statistical taggers may perform poorly with medical documents. This is supported by our finding that the distribution of POS in three corpora was quite different. We found less syntactic variation in both surgical pathology reports and MEDLINE abstracts when compared with WSJ documents. MEDLINE abstracts and SPRs have higher frequencies of nouns when compared to Wall Street Journal articles. Therefore, the POS transitional information in medical documents is likely to be different as well as the features of each word.

We found that a general English trained model trained on the Wall Street Journal did not perform well on SPRs. This reproduces Coden's observation on both clinical notes and Pub-Med abstracts.[3] We also observed that use of an 800 term domain lexicon in conjunction with general English achieved only a 5% increase in accuracy from 79% to 84%. This is comparable to Coden's findings which showed only a 2% increase in accuracy in POS tagging over general English.[3] The resulting performance is inadequate to support further NLP components such as parsers that rely on POS information. Domain lexicons suffer from a lack of contextual information which is important to enhance per-

Table 4 ■ Partial Confusion Matrix Showing Distribution of Most Frequent POS Tagging Errors by ME Tagger Trained with WSJ and DSC (Data Set 3)

Tagging Error	Reference Standard	ME Tagger trained with WSJ + DSC										
		NN	JJ	NNP	VBD	VBN	CD	NNS	RB	VB	DT	% of Total
JJ		19% (49)	—	2% (4)	0% (1)	1% (2)	—	—	—	1% (3)	—	24% (64)
LS		6% (17)	0% (1)	8% (20)	—	—	3% (7)	—	—	—	1% (2)	19% (49)
NN		—	11% (28)	1% (3)	0% (1)	—	—	—	—	—	—	13% (34)
VBN		—	—	—	8% (21)	—	—	—	—	—	—	8% (22)
VBD		—	—	—	—	8% (20)	—	—	—	—	—	8% (20)
NNP		3% (7)	3% (9)	—	—	—	—	0% (1)	—	—	—	7% (18)
NNS		5% (13)	0% (1)	—	—	—	—	—	—	—	—	6% (15)
CC		—	—	—	—	—	—	—	1% (3)	—	—	2% (6)
IN		—	—	—	—	—	—	—	1% (2)	—	—	2% (5)
RB		—	1% (3)	—	—	—	—	—	—	—	—	2% (5)
% of Total		36% (94)	18% (47)	11% (28)	9% (23)	8% (22)	3% (8)	3% (7)	2% (6)	2% (5)	0% (3)	90% (238)

JJ = adjective; LS = list item marker; NN = noun, singular or mass; VBN = verb, past participle; VBD = verb, past tense; NNP = noun, proper, singular; NNS = noun, plural; CC = conjunction, coordinating; IN = preposition or conjunction, subordinating; RB = adverb; CD = numeral, cardinal; VB = verb, base form; DT = determiner.
Number of errors shown in percentage of total errors, with counts in parentheses, for most frequent errors.

Table 5 ■ Partial Confusion Matrix Showing Distribution of Most Frequent POS Tagging Errors by ME Tagger Trained with WSJ + H1 Selected Data from DSC (Data Set 5)

Tagging Error	Reference Standard	ME Tagger retrained with WSJ + H1 selected data										% of Total
		NN	JJ	NNP	VBN	VBD	NNS	RB	CD	VBG	IN	
JJ		18% (56)	—	2% (7)	3% (8)	1% (4)	—	1% (4)	—	2% (6)	—	28% (89)
LS		5% (17)	1% (4)	9% (30)	—	—	—	—	2% (7)	—	—	19% (60)
NN		—	15% (47)	1% (3)	—	0% (1)	1% (4)	—	—	—	—	18% (58)
VBD		—	—	—	6% (20)	—	—	—	—	—	—	6% (20)
VBN		—	0% (1)	—	—	5% (16)	—	—	—	—	—	5% (17)
NNS		3% (8)	1% (3)	1% (3)	—	—	—	—	—	—	—	5% (15)
NNP		2% (5)	2% (7)	—	—	—	1% (2)	—	—	—	—	4% (14)
IN		—	—	—	—	—	0% (1)	1% (2)	—	—	—	3% (9)
CC		—	—	—	—	—	—	1% (3)	—	—	1% (2)	2% (6)
CD		1% (2)	1% (2)	—	0% (1)	—	—	—	—	—	—	2% (5)
% of Total		30% (96)	22% (69)	14% (44)	9% (29)	7% (21)	3% (10)	3% (10)	2% (7)	2% (6)	1% (4)	92% (293)

JJ = adjective; LS = list item marker; NN = noun, singular or mass; VBD = verb, past tense; VBN = verb, past participle; NNS = noun, plural; NNP = noun, proper, singular; IN = preposition or conjunction, subordinating; CC = conjunction, coordinating; CD = numeral, cardinal; RB = adverb; VBG = verb, present participle or gerund.
Number of errors shown in percentage of total errors, with counts in parentheses, for most frequent errors.

formance of statistical taggers. Overall, our data support the contention by Coden and colleagues[3] that domain specific training corpora are required to achieve high accuracy for POS taggers in medical domains.

If domain lexicons are inadequate, and large training sets are impractical to develop, how can medical NLP researchers develop accurate domain-specific statistical taggers? We hypothesized that sample selection might provide a method for development of small but highly efficient training sets. If true, researchers could develop smaller training sets which would be easier, faster, and cheaper to develop but might achieve nearly the same result as larger, more general training sets.

Table 6 ■ Comparison of Errors Assigning POS for Tagger Trained with Three Different Corpora

Reference tag → Error	ME tagger trained with WSJ only	ME tagger trained with WSJ + DSC	ME tagger trained with WSJ + H1 selected Data
CC → RB	3	3	3
CD → JJ	0	0	2
CD → NN	1	0	2
CD → NNP	21	0	0
IN → NNP	10	0	0
IN → RB	3	2	2
JJ → NN	77	49	56
JJ → NNP	271	4	7
LS → NNP	70	20	30
NN → JJ	23	28	47
NN → NNP	310	3	3
NNP → JJ	2	9	7
NNS → NN	89	13	8
RB → JJ	7	3	2
VBD → VBN	57	20	20
VBN → NNP	12	0	0
VBN → VBD	10	21	16
VBZ → NNS	14	3	3

CC = conjunction, coordinating; RB = adverb; CD = numeral, cardinal; JJ = adjective; NN = noun, singular or mass; NNP = noun, proper, singular; IN = preposition or conjunction, subordinating; LS = list item marker; NNS = noun, plural; VBD = verb, past tense; VBN = verb, past participle; VBG = verb, present participle or gerund.

Our data showed that heuristics based on comparative frequencies provides a powerful method for selecting a smaller training set. The highest gain in accuracy was obtained when we selected sentences that contained the most frequent unknown words. The accuracy of POS tagging on surgical pathology reports was boosted substantially to another 8.7% over domain lexicons adaptation (from 84% to 92.7%). The H1 selected sentences provided the same frequency information available in a domain lexicon but also included the contextual information that a domain lexicon does not have. This result seems especially promising since the upper bound accuracy was 93.9% when the entire domain corpus was used for training. In our study, we only needed to annotate approximately 665 sentences and 5,114 terms. This number can be decreased by a further 50% based on findings of the learning curve study. Taken together, these results show that an 84% total decrease in sample size can be achieved without any sacrifice in performance.

Heuristic selection produced a distribution of errors in POS assignment that was very similar to the distribution obtained with the entire domain specific corpus, which strengthens our conclusion that heuristic sample selection can be used with few disadvantages. Error analysis revealed that many of the remaining errors produced by the ME tagger could be corrected by pre-processing of the data—for example list item markers may be correctly identified if outlines in clinical reports can be identified prior to POS tagging.

Future Work

In this study we demonstrated the potential of heuristic sample selection to minimize training set requirements for lexical annotation of medical documents. The simple heuristics we used were highly effective. We are interested in evaluating several other selection heuristics for their relative effect on performance. Additionally, we intend to incorporate these algorithms into the open source GATE annotation

environment as processing resources for medical corpora development. These tools may be of benefit for corpus development of many different NLP components in a variety of health-related domains.

Conclusion

An ME tagger retrained with a small domain corpus created with heuristic sample selection performs better than the native ME tagger trained on English, MedPost POS tagger trained on MEDLINE abstracts, and the ME tagger retrained with a domain specific lexicon. Sample selection permits a roughly 84% decrease in size of the annotated sample set, with no decrease in performance of the retrained tagger We conclude that heuristic sample selection based on frequency and uncertainty, provides a powerful method for decreasing the effort and time required to develop accurate statistically-based POS taggers for medical NLP.

References ■

1. Marcus M, Marcinkiewicz M, Santorini B. Building a large annotated corpus of English: the penn treebank. Comp Ling 1993;19(2):313–30.
2. Gupta D, Saul M, Gilbertson J. Evaluation of a DeIdentification (deID) Software Engine to Share Pathology Reports and Clinical Documents for Research. Am J. Clin Pathol 2004;121:176–86.
3. Coden AR, Pakhomov SV, Ando RK, Duffy PH, Chute CG. Domain-specific language models and lexicons for tagging. J Biomed Inform 2005;38(6):422–30.
4. Taira R, Soderland S, Jakobovits R. Automatic structuring of radiology free-text reports. Radiol 2001;21(1):237–45.
5. Schadow G, McDonald C. Extracting structured information from free text pathology reports. Proc AMIA Symp 2003:584–8.
6. Stetson P, Johnson S, Scotch M, Hripcsak G. Sublanguage of Cross-coverage. Proc AMIA Symp 2002:742–6.
7. Chapman W, Bridewell W, Hanbury P, Cooper G, Buchanan B. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp 2001:105–9.
8. Campbell DA, Johnson SB. Comparing syntactic complexity in medical and non-medical corpora. Proc AMIA Symp 2001:90–4.
9. Ceusters W, Buekens F, De Moor G, Waagmeester A. The distinction between linguistic and conceptural semantic in medical terminology and its implication for NLP-based knowledge acquisition. Proc IMIA WG6 Conf Nat Lang Med Conc Rep. Jacksonville 19–22/01/97:71–80.
10. Grover C, Lascarides A. A comparison of parsing technologies for the biomedical domain. Nat Lang Eng 2005;11(1):27–65.
11. Baud R, Lovis C, Rassinoux AM, Michel PA, Scherrer JR. Automatic extraction of linguistic knowledge from an international classification. Proc MEDINFO'98, Seoul, Korea 1998:581–5.
12. Smith L, Rindflesh T, Wilbur WJ. MedPost: a part of speech tagger for biomedical text. Bioinform J 2004;1(1):1–2.
13. Weischedel R, Meteer M, Schwartz R, Ramshaw L, Palmucci J. Coping with Ambiguity and Unknown Words through Probalistic Models. Comp Ling 1993;19(2):359–82.
14. Jelinek F, Lafferty J, Magerman D, Mercer R, Ratnaparkhi A, Roukos S. Decision Tree Parsing using a Hidden Derivational Model. In Proc Human Lang Technol Workshop (ARP, 1994) 1994:272–7.
15. Magerman DM. Statistical Decision-Tree Models for Parsing. In Proc 33rd Ann Meeting ACL. 1995.
16. Toutanova K, Klein D, Manning CD, Singer Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. Proc HLT-NCAL 2003:173–80.
17. Brill E. Some Advances in Transformation-Based Part of Speech Tagging. In Proc Twelfth Nat Conf Artif Intell 1994;1:722–7.
18. Ratnaparkhi A. A Maximum Entropy Model for Part-Of-Speech Tagging. In Proc Empir Meth Nat Lang Proc Conf, May 17–18 1996.
19. The Penn Treebank Project. Available at: http://www.cis.upenn.edu/~treebank/. Accessed on July 27, 2007.
20. Divita G, Browne A, Loane R. dTagger: a POS Tagger. Proc AMIA Symp 2006:200–3.
21. Tsutuoka Y, Tateishi Y, Kim J, Ohta T, McNaught J, Ananiadou S, et al. Developing a Robust Part-of-Speech Tagger for Biomedical Text. Advances in Informatics—10th Panhellenic Conf Inform, LNCS 2005:382–92.
22. Tateisi Y, Tsuruoka Y, Tsujii J. Subdomain adaptation of a POS tagger with a small corpus. Proc BioNLP Workshop 2006:136–7.
23. Fujii A, Inui K, Tokunaga T, Tanaka H. Selective Sampling for Example-based word sense disambiguation. Comput Ling 1998; 24(4):573–98.
24. Hwa R. Sample Selection for Statistical Parsing. Comput Ling 2002;30(3).
25. Maximum Entropy Part of Speech Tagger. Software may be downloaded from Download Site (July/2007): http://www.cogsci.ed.ac.uk/~jamesc/taggers/MXPOST.html. Accessed on July 27, 2007.
26. SPECIALIST Lexicon and tools. Available at: http://specialist.nlm.nih.gov/. Accessed on July 27, 2007.
27. Santorini B. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania. 1990.
28. Carletta J. Assessing agreement on classification tasks: The Kappa statistic. Comput Ling 1996;22(2):249–54.