Development of a Pediatric Text-Corpus for Part-of-Speech Tagging

John Pestian[1], Łukasz Itert[1,2], and Włodzisław Duch[2,3]

[1] Department of Pediatric Informatics, Children's Hospital Research Foundation, Cincinnati, OH, USA
[2] Department of Informatics, Nicholas Copernicus University, Torun, Poland
[3] School of Computer Engineering, Nanyang Technological University, Singapore

Abstract. Most efforts in natural language processing (NLP) have been devoted to understanding general domain data. Special domains, such as pediatric medicine, pose some unique problems and challenges. While many common sense corporas and lexicons have been created we know of none directly related to pediatric medicine. This article presents the status of an ongoing project to create a large corpus and lexicon for use by part-of-speech tagger and other NLP research tools, aimed at developing new methods in sciences related to medical domains. Experiments with automatic tagging set the limit of attainable accuracy at 92-93% on this type of texts.

1 Introduction

1.1 CCHMC project

The project, initiated at the Cincinnati Children's Hospital Medical Center (CCHMC), aims at developing an intelligent system that can support clinical-genomic research. Given the significant increase in the volume of clinical, genetic, and genomic data, it is essential that intelligent systems be developed that can aid the care giver in the diagnosis, selection and optimization of patients treatment. The CCHMC data repository, called Discovery System, consists of 5 Terabytes of the medical and genomic annotations that include data containing nurses and surgical notes, discharge summaries, information about symptoms, procedures, findings, and therapeutic response, genetic specimens, and so forth.

Analysis of such complex data requires merging of Natural Language Processing (NLP) and Computational Intelligence (CI) methods. The entire project is divided into the following steps:

- Preprocessing of free-text.
- Assignment of part-of-speech (POS) to preprocessed text.
- Assignment of words tagged by POS to Unified Medical Language System (UMLS) concept space.
- Applying existing CI methods or developing new ones to achieve the planned goals.

Stage one, preprocessing free-text, is currently underway. Stage two is based on the UMLS [1], which is the largest medical ontology created so far. This knowledge and understanding of the medical domain are provided by ontology. Authors of UMLS describe their goals and intentions as [2]: *to aid the development of systems that help health professionals and researchers retrieve and integrate electronic biomedical information from a variety of sources*. UMLS consists of three parts:

- The Metathesaurus contains semantic information about biomedical concepts, their various names, and relationships among them.
- The Semantic Network provides a consistent categorization of all concepts represented in the Metathesaurus.
- The Specialist lexicon contains syntactic information about biomedical terms.

The Specialist lexicon is used at the first stage and plays the central role as the main, elementary lexicon. NIH also provides the software package called MetaMap Transfer (MMTx) [3] which: *maps arbitrary text to concepts in the UMLS Metathesaurus; or, equivalently, it discovers Metathesaurus concepts in text*. To work properly, MMTx requires a Part Of Speech (POS) tagging system otherwise MMTx is unable to properly classify text.

1.2 Remarks about medical data

A variety of taggers tend to give the best possible accuracy of classifying tokens that are not represented by specific domain. These taggers require an underlying body of knowledge, or corpus, that act as a reference set for the classification problem. The most widely used dataset is Penn Treebank Wall Street Journal corpus [4]. Although it is very large (over 1 million classified words) and well maintained, it derives from general domain making it unsuitable for classification of medical texts. Medical texts require unique treatment because:

- They describe a specific domain which is not understood by the laypersons.
- They are dense with medical terms which are specific and designed to use only in one particular branch of medicine.
- Typical grammar structures can be violated without losing context.

The medical data involved in this project has a few additional features that make it challenging to tag because the text comes from both dictated and manually written down annotations. These additional challenges include:

- Frequent misspelling and typing errors.
- Frequent punctuation errors with commas and periods that result in sentence classification errors.

- Large number of ambiguous abbreviations. For example, OR can represent the token "operating room" as well as the conjunction "or"; BE may represent "barium enema" or "base excess".

- There are frequent one-character symbols like

  `#,",~,-,<,>`

  which carry information and should not be just disregarded.
- Each subspeciality of medicine uses different standards and formats for storing records, dates, etc.
- Frequent occurrence of multi-words.
- Patients' personal data found within the text needs to be de-identified.

In general English language multi-words, such as "ice cream", "data base" or "follow-up", composed of several words referring together to one concept or item, are quite rare, but in medical domain they may be common. The number of words in a string may vary; words may be separated by a space or a hyphen. Although multi-words consist of a few parts, only one part of speech should be assigned to them. In [5] a brief comparison of several sources of English is presented; considering the percentage of strings having more than three words, dramatic difference is noticed: for Webster's Dictionary (the largest source of general American English) it's 0.003%, but for UMLS Metathesaurus it's in fact more than half, 53%.

Thus, all these differences suggest that the utility of a general knowledge corpus within the medical domain is limited.

2 Methods

At this stage our goal is to develop a pediatric medical corpus that can be used to classify pediatric medical text by parts of speech. This should enable accurate mapping of the text on the UMLS ontology using MMTx software. The methods used to achieve this goal include:

1. Preprocessing raw text to increase the reliability and validity of the text by making it more readable, regular, standard and anonymous without losing contextual value.
2. Tagging the processed text by correct part-of-speech using human expert linguists.
3. Using the hand-tagged text as a training set for part-of-speech tagging software.
4. Continue adding training sets until a 95% accuracy is achieved using a 10x cross-validation.

Further discussion of steps one and three is given below.

Data Source: Our primary data is the set of clinical annotations from different branches of medicine (surgery, radiology, etc) that were collected during 2002. These data come from our Discovery System, a clinical-genomic data repository [6]. In total there are over 20 million tokens, grouped into over 1 million of sentences. Preprocessing the raw text requires the following main steps:

1. Classify raw text into sentences and words.
2. Search for medical and common sense abbreviations and symbols in text and replace them with proper meanings.
3. Search for ambiguous abbreviations and replace them with suitable meaning based on surrounding text.
4. Proper processing of multi-words.
5. Making confidential data harmless.

Creating the tagging reference set: From the data set, random sentences are selected and given to a linguist group to determine the appropriate part of speech. As these tagged data return they are added to the main training set.

2.1 Tagging process

Tagging systems can be classified into several groups. They can be divided into supervised vs. unsupervised systems, or statistical vs. based on maximum entropy principles. On the other hand we can distinguish N-gram tagging, transformation based, or rule tagging [7]. All these systems achieve very good results when applied to the Wall Street Journal data, giving accuracies between 95-98% on the separate test set.

Focusing on supervised systems (which seem to be more universal and widely used), one can easily notice that all of them need a lexicon (providing the list of words, each with the part of speech tag assigned to it) and the training set (giving information about correlation between part of speech names within a sentence).

In the present project TreeTagger [8] has been selected, a supervised algorithm developed by Helmut Schmid. It is based on the slightly modified ID3 decision tree and adapted to work with textual data. TreeTagger is very fast and has great support for misspelled words as well as words non-existing in the lexicon.

Validation Process: Since the development of the training set continues, we are using N-fold Cross Validation (CV) process for determining overall accuracy, where N=10. Processing of each fold consists of the following steps:

1. From randomized main tagged corpus select 1/N-th number of sentences. Remove all parts of speech tags to create the Independent Set (IS).
2. Use the remaining data as the training set.

3. Train the TreeTagger on this data.
4. Upon completion use the IS for tests to determine automatic tagging accuracy.

The total accuracy (presented in charts) is determined by computing the average of three 10 CV results. To estimate the accuracy on the training set, the same set for both training and testing is used.

3 Results

Our investigation is aimed at creating both the lexicon and the training set that are useful for pediatric medical data. The Specialist lexicon from UMLS was used as a base, extended by 265.000 additional unique common-sense entries that were available to the lexicon. The question of an optimal training set size still remains unanswered. One could say that the bigger the set is the better, but this approach is not the most efficient use of time and funds (it involves manual tagging). The size of the training set used in experiments has been increased by 5,000 tokens, and then the tagger was retrained. This systematic approach enables determination of the number of tokens that already give marginal gain in accuracy.

The change of accuracy as a function of the training set size is presented in Figure 1. Currently our training set size has 390.000 words, however we stopped doing regular computations at 210.000 words and begun to process more data at once. By now, the average accuracy from 10CV is 91.5%. This result includes contribution from the support for multi-words we have added to TreeTagger (taking the advantage of the UMLS Lexicon as a base). Having done this, the tagging accuracy increased by 0.5%, but the real benefit is expected to be seen only while discovering the UMLS concepts in the text.

To estimate the upper limit of the growth of the accuracy on the test set, all calculations were repeated for the full training set. TreeTagger was learning on the full training set and then the same set was used to make the predictions. Comparison of both, training and test accuracies are shown in Figure 2. The training accuracy for the full set is currently 93.0%.

The test accuracy grows very slowly and the convergence seems to be exponentially slow. To ensure that this assumption is correct, NSP [9] package has been used to count the number of unique trigrams (groups of three words in a row) in the training text. TreeTagger uses trigrams during its learning process. Counting the number of unique trigrams may be important, believing that if something occurs only once in the 20 million words set, it just has to be irrelevant. Results from these calculations are presented in Figure 3.

In the 210.000 word there are about 43% of unique trigrams, in the entire 20 million word set there still 16% of them. The decrease of the percent of number of unique trigrams seems to be exponential, making the increase of the number of trigrams approximately linear for large number of words. Since

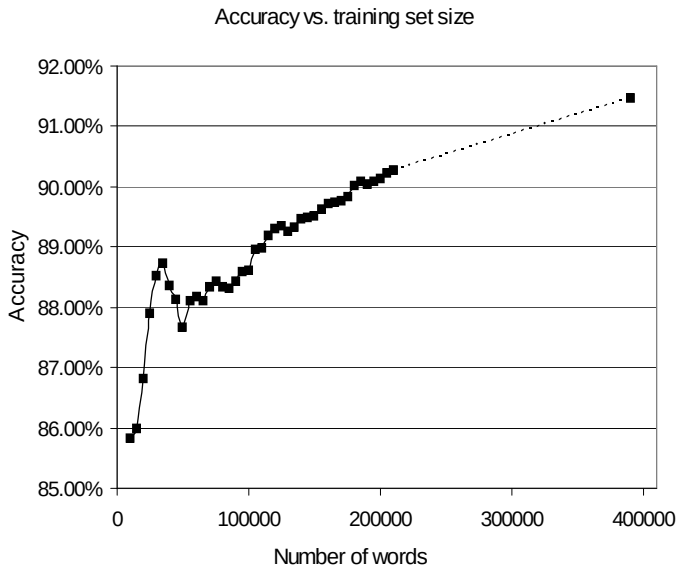

Fig. 1. The tagging accuracy vs. the training set size. All results are from the 10-CV tests. Black squares show points where actual calculations were done.

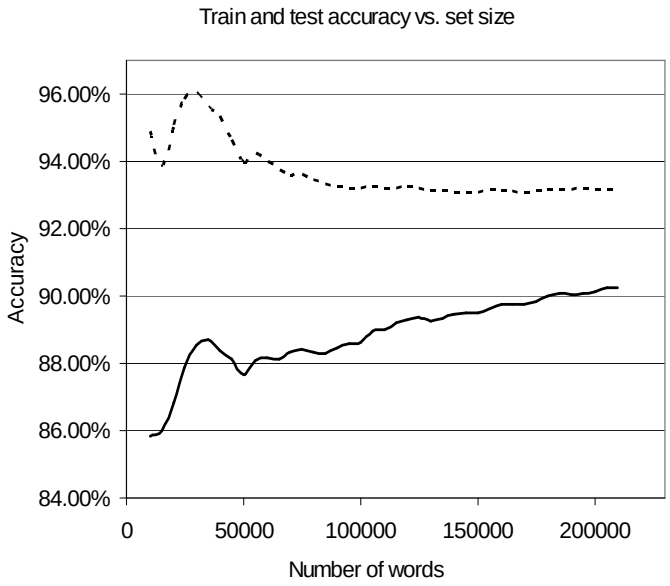

Fig. 2. Accuracies on the training (the dashed line) and the test set (the solid line) for 215000 words.

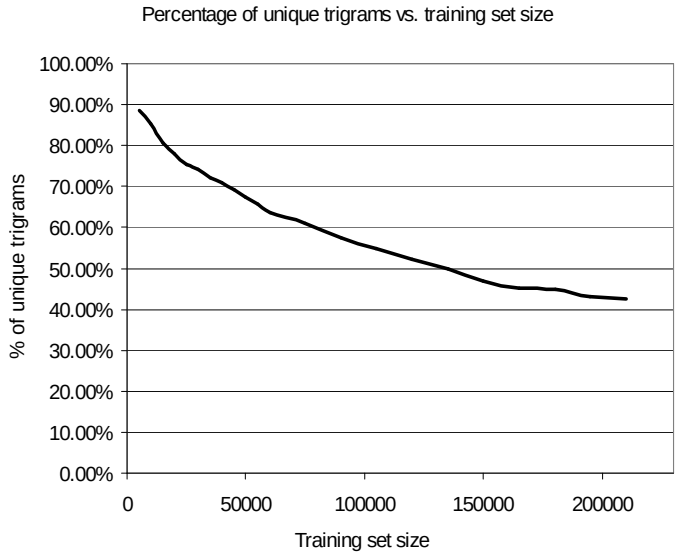

Fig. 3. The percentage of unique trigrams vs. the training set size.

generalization from single trigrams is not possible the test accuracy will also saturate exponentially.

4 Conclusions

The project is still in its early stages and the training set is still expanding. Although a significant improvement have been noticed, the results show that the learning process discussed here is likely to converge slowly. Therefore, a large training set may be necessary for obtaining the high accuracy. Tests made on the training set show that the upper limit for the top test accuracy ranges between 92-93%. Although humans use more complex interpretation mechanisms the information available in text alone will not be sufficient to reach 100% accuracy. It would be very interesting, but also quite difficult, to find out the accuracy limit for humans on the same task. If the answer is significantly higher than 93% then adding more knowledge-based information to the tagging process should enable to remove some of the errors. However, creation of taggers capable that could use non-statistical information remains a challenge.

References

1. http://www.nlm.nih.gov/research/umls
2. UMLS Knowledge Sources, 13th Edition January Release.
3. http://mmtx.nlm.nih.gov
4. M. Marcus, B. Santorini, and M. Marcinkiewicz, Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19: 313-330, 1993.
5. A.T. McCray, O. Bodenreider, J.D. Malley, A.C. Browne, Evaluating UMLS strings for Natural Language Processing. Proceedings of AMIA Annual Symposium 2001; 448-452.
6. J. Pestian, B. Aronow, K. Davis, Design and Data Collection in the Discovery System. International Conference on Mathematics and Engineering Techniques in Medicine and Biological Science, 2002.
7. E. Brill, J. Wu, Classifier Combination For Improved Lexical Disambiguation, COLING/ACL, 1998.
8. H. Schmid, Probabilistic Part–of–Speech Tagging Using Decision Trees. Proceedings of the Conference on New Methods in Language Processing, 1994.
9. S. Banerjee, T. Pedersen, The Design, Implementation and Use of the Ngram Statistics Package. In: Pro. 4-th Int. Conf. on Intelligent Text Processing and Computational Linguistics, Mexico City, pp. 370-381, 2003.