Text Mining for Medical Documents Using a Hidden Markov Model

Hyeju Jang[1], Sa Kwang Song[2], and Sung Hyon Myaeng[1]

[1] Department of Computer Science, Information and Communications University, Daejeon, Korea
{hjjang, myaeng}@icu.ac.kr
[2] Electronics and Telecommunications Research Institute, Daejeon, Korea
smallj@etri.re.kr

Abstract. We propose a semantic tagger that provides high level concept information for phrases in clinical documents. It delineates such information from the statements written by doctors in patient records. The tagging, based on Hidden Markov Model (HMM), is performed on the documents that have been tagged with Unified Medical Language System (UMLS), Part-of-Speech (POS), and abbreviation tags. The result can be used to extract clinical knowledge that can support decision making or quality assurance of medical treatment.

1 Introduction

Patient records written by doctors are invaluable information especially in areas where experiences have great consequences. If doctors find useful information they need from patients' records readily, they can use it to deal with problems and treatments of current patients. That is, it can provide a support for medical decision making or for quality assurance of medical treatment.

In the treatment of chronic diseases, for example, the past records on the symptoms, therapies, or performances a patient has shown assist doctors to get a better understanding of different ways of controlling a disease of the current patient. As a result, they help their decisions for the direction of the next treatment.

Moreover, hospitals where medical records are kept in the computers are increasing nowadays. The growing availability of medical documents in a machine-readable form makes it possible to utilize the large quantity of medical information with linguistically and statistically motivated tools. Implicit knowledge embedded in a large medical corpus can be extracted by an automated means.

This paper describes a tagging system that yields high-level semantic tags for clinical documents in a medical information tracking system. The tags in this system are categories of information that phrases of medical records contain, such as *symptom, therapy,* and *performance*. They will allow the tracking system to retrieve past cases doctors want to know about a certain therapeutic method, for example. The tagging system uses existing medical terminological resources, and probabilistic Hidden Markov Models [1] for semantic annotation.

H.T. Ng et al. (Eds.): AIRS 2006, LNCS 4182, pp. 553 – 559, 2006.

The contributions of this research can be summarized in three aspects. First, from a practical point of view, it widens the possibility of helping doctors with the experiences and knowledge embedded in the past patient records. Second, from a technical point of view, it attempts to annotate clinical text on phrases semantically rather than syntactically, which are at higher level granularity than words that have been the target for most tagging work. Finally, it uses a special method to guess unknown phrases that don't appear in the training corpus for the robust tagging.

2 Related Works

The popular and conventional approach of part-of-speech (POS) tagging systems is to use a HMM model so as to find a most proper tag [2]. Some systems use a HMM with additional features. Julian Kupiec [3] and Dong Cutting et al. [4] described POS tagging systems, which have the concept of ambiguity class and equivalence class, respectively. Our system also adopted the equivalence class concept which group words into equivalence classes.

Tagging systems in the medical field have focused on the lexical level of syntactic and semantic tagging. Patrick Ruch [5] and Stephen B. Johnson [6] performed semantic tagging on terms lexically using the Unified Medical Language System (UMLS). On the other hand, Udo Hahn et al. [7] and Hans Paulussen [8] built POS taggers which categorized words syntactically.

There also have been the systems which extract information from the medical narratives [9, 10, 11]. Friedman [9, 10] defined six format types that characterize much of the information in the medical history sublanguage.

3 Methodology

The purpose of the tagging system is to annotate the clinical documents with semantic tags that can be used by a tracking system whose goal is to provide useful information to doctors. Our work is based on the list of questions doctors are interested in getting answers for, which was provided by Seoul National University Hospital (SNUH). Among them, we focused on the two questions: 'How can X be used in the treatment of Y?' and 'What are the performance characteristics of X in the setting of Y?' where X and Y can be substituted by {Medical Device, Biomedical or Dental Material, Food, Therapeutic or Preventive Procedure} and {Finding, Sign or Symptom, Disease or Syndrome}, respectively. Our tagging system assigns semantic tags to appropriate phrases so that the tracking system can answer those questions.

The semantic tags were chosen to answer the questions from the doctors in SNUH. While there are many interesting questions and therefore many tags to be used ultimately by a tracking system, we chose Symptom, Therapy, and Performance as the Target Semantic Tags (TST) for the current research. *Symptom* describes the state of a patient whereas *Therapy* means everything a medical expert performs for the patient, such as injection, operation, and examination. *Performance* means the effect or the result of a therapy and includes the results of some examinations or the change of a patient's status (e.g. getting better or getting worse).

TST in this research distinguish the tagging system unique because they represent higher level concepts. Unlike part-of-speech (POS) or UMLS semantic categories of a term, TST can be utilized by the application systems directly. In fact, TST was chosen for a particular application system in the first place. The categories of TST should be changed depending on the purpose of the application system, but the method we propose can be used in the same manner with an appropriate training corpus.

There can be different ways of assigning semantic tags to phrases. Our work is based on an observation that there is a specific sequence when people record something. For example, a description on a cause is followed by that of an effect. Events are usually described in their temporal order. We assumed that the narrative data in CDA documents has implicit rules about sequences.

In order to model the sequential aspect of the clinical documents, we opted for Hidden Markov Model (HMM). Unfortunately, we cannot fully use the grammar rules in our research because our corpus includes Korean and English words mixed. But with the idea that people tend to write things in a certain sequence, we chose to use HMM.

The system architecture for the semantic tagger using HMM is shown in Fig. 1. It is divided into two stages: training and tagging.

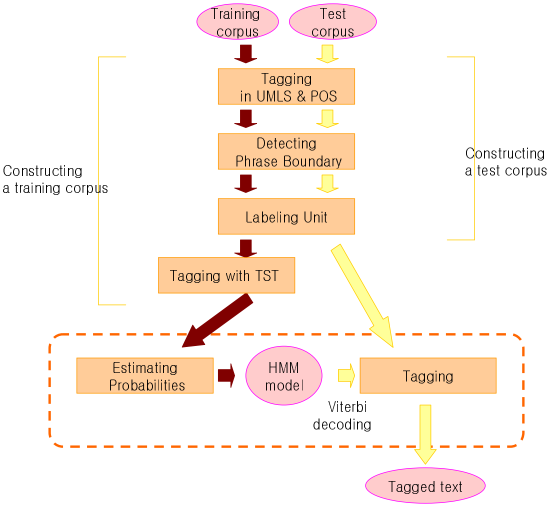


Fig. 1. The system architecture for the TST tagger

3.1 Common Part

1) Tagging in UMLS & POS: The corpus is first processed with UMLS tagging and POS tagging. The former is for classifying medical terms in their semantics whereas the latter is for understanding the syntactic role of words. Abbreviations in the corpus are processed based on the research in the same project.

2) Detecting Phrase Boundary: This is important because symptom, therapy, or performance in TST is described with a phrase or a whole sentence, not a word. This task is not as simple as that for other types of text since doctors usually don't write grammatically correct sentences. In addition, periods are used not only for indication of the end of a sentence but also for abbreviations, dates, floating point numbers and so on.

A phrase is defined to be a unit that ends with a predicate (i.e. a verb ending in Korean) and include a subject with some intervening words like function words and adverbs. Since doctors tend to write a subject like a lab test or medication in English and a predicate in English, a phrase tends to consist of both English and Korean words.

3) Labeling Phrase Units with Equivalence Class: Since there are many words occurring only once in the corpus, we place words into equivalence classes so that class labels are used in HMM (see Table 1 for the equivalence classes). Words are grouped into equivalence classes, and a phrase is expressed with the set of equivalence classes it contains. Fig. 2 shows how a phrase is transformed into an observance expressed with equivalence classes.

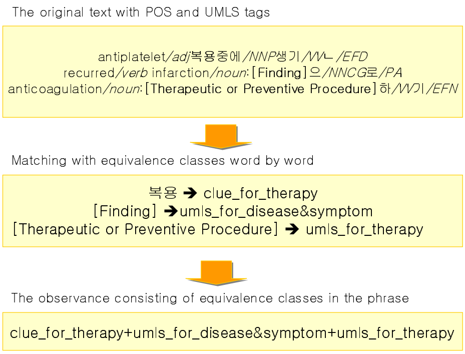
Table 1. Equivalence Classes for Words

UMLS tag for cause	Biomedical or Dental Material, Food
UMLS tag for disease or symptom	Finding, Sign or Symptom, Disease or Syndrome, Neoplastic Process
UMLS tag for therapy	Diagnostic Procedure, Food, Medical Device, Therapeutic or Preventive Procedure
Clue word for therapy	처방(prescription), 복용(administer medicine), 시행(operation), 후(after), 이후(later), 사용(use), 증량(increase), 수술(surgery), 중단(discontinue)
Clue word for symptom	발열(having fever), 관찰(observe)
Clue word for performance	호전(improvement), 감소(decrease), 상승(rise), 정상(normal), 발생(occurrence), 변화(change)
unknown	neither clue word nor UMLS tag

3.2 Training Part

1) Tagging with Target Semantic Tags (TST): For the training corpus, the tagging is done manually.

2) Estimating Probabilities: There are two training methods used in current tagging systems. One is to use the Baum-Welch algorithm [16] to train the probabilities, which does not require a tagged training corpus. Another advantage is that for different TST, the only thing we should do is just replace the corpus. The other method is to use tagged training data. This method counts frequencies of words/phrases/tags to estimate the probabilities required for a HMM model. The disadvantage of this method is that it needs a tagged training corpus whose quantity is enough to estimate the probabilities. Building a training corpus is a time-consuming and labor intensive work. Despite this disadvantage, we choose the second method because its accuracy is much higher than that of the Baum-Welch method. David Elworthy [13] and Bernard Merialdo [14] compared tagging performance using the Baum-Welch algorithm against the one using the tagged-training data, proving that using training data is much more effective.


The original text with POS and UMLS tags
antiplatelet*/adj*복용중에*/NNP*생기*/VV*ㄴ*/EFD*
recurred*/verb* infarction*/noun*: [Finding] 으*/NNCG*로*/PA*
anticoagulation*/noun*: [Therapeutic or Preventive Procedure] 하*/VV*기*/EFN*
Matching with equivalence classes word by word
복용 ➔ clue_for_therapy
[Finding] ➔umls_for_disease&symptom
[Therapeutic or Preventive Procedure] ➔ umls_for_therapy
The observance consisting of equivalence classes in the phrase
clue_for_therapy+umls_for_disease&symptom+umls_for_therapy

* Italic fonts denote POS tags.
* Brackets denote UMLS tags.
** "복용중에" means "in the middle of taking medicine"
** "생기" means "happen"
** "하" means "do"
** "으", and "로" are a case marker in Korean language.

Fig. 2. The observance of a phrase with equivalence classes

3.3 Tagging Part

1) Tagging: The system finds a most probable tag sequence using the Viterbi algorithm [15] using the HMM model constructed in the training stage.

2) Tagging of unknown phrase units: phrases appearing in the test corpus are categorized into largely two groups. The first group is for a phrase with no component word known to the system and hence transformed to an equivalence class label. There is no clue in the phrase that can be used in predicting its meaning. Since the whole phrase is labeled as *unknown*, not a class label, its statistics can be gathered from the training corpus that contains many unknown phrases. The other group is for the unknown phrases that have some clues with the words comprising the phrase unit, which have their class labels. The reason why they are called *unknown* is because the particular combination of the class labels corresponding to the phrase is not simply available in the training corpus. We call such a clue combination, not sequence, a *pattern*. The probability of an unknown phrase can be estimated with the equivalence class labels although the unit itself is unknown (see Fig. 2 for an example).

When an unknown pattern appears as an observance, it is compared against the existing patterns so that the best pattern can be found, to which the unknown pattern can be transformed. That is, an unknown pattern is regarded as the best matching pattern. The pattern that matches best with the unknown pattern is chosen and its probability is the same as that of the selected pattern. The probability of that unknown pattern of observance is calculated using the probability of the most similar pattern. When more than one pattern is most similar, the probability of the unknown pattern becomes the average of the most similar patterns.

4 Experiments and Results

4.1 Data

The Clinical Document Architecture (CDA) provides a model for clinical documents such as discharge summaries and progress notes. It is an HL7 (Healthcare Level 7) standard for the representation and machine processing of clinical documents in a way that makes the documents both human readable and machine processable and guarantees preservation of the content by using the eXtensible Markup Language (XML) standard. It is a useful and intuitive approach to management of documents which make up a large part of the clinical information processing area [12].

We picked 300 sections of “progress after hospital stay” from the CDA documents as the target corpus provided by SNUH for research purposes.

The training corpus consists of 200 “progress after hospital stay” sections containing 1187 meaningful phrases that should be tagged. The test corpus is 100 sections with 601 phrases.

4.2 Performance

The level of accuracy of our system is calculated as the number of correct tags per the total number of tags. Fig.3 shows the comparison of the basic model with and without unknown phrase processing. Although the result of the system is not as good as expected, it is promising and undergoing further improvement. We suggest the direction of modification below.

- Increase the number of different equivalence classes. The number of tags corresponding to the equivalence classes is so small at this point that the transition probabilities are not very meaningful.
- Find better initial probabilities of a HMM model.
- Improve the unknown phrase guessing method.
- Get as much tagged text as we can afford.

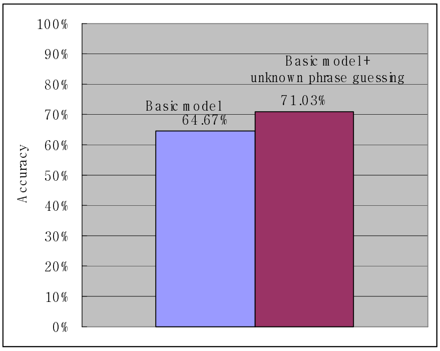


Fig. 3. The results of a basic HMM model and a HMM model with an unknown phrase guessing module

5 Conclusion

We showed a semantic tagger for medical documents using a HMM model. For future work, we are going to utilize symbols and numeric expressions to represent phrases better and to find a better way for matching equivalence classes. Moreover, we will design and compare against other methods such as Markov random fields, SVM, and so on.

Acknowledgement

This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Korea.

References

1. L.R.Rabiner et al., “An Introduction to Hidden Markov Models”, IEEE ASSP Magazine, 1986
2. Linda Van Guilder, “Automated Part of Speech Tagging:A Brief Overview”, Handout for LING361, 1995
3. Julian Kupiec, “Robust part-of-speech tagging using a hidden Markov model”, Computer Speech and Language, pp. 225–242, 1992.
4. Doug Cutting et al., “A Practical Part-of-Speech Tagger”, In Proceedings of the 3rd ACL, pp.133–140, 1992
5. Patrick Ruch, “MEDTAG: Tag-like Semantics for Medical Document Indexing”, In Proceedings of AMIA’99, pp.35–42
6. Stephen B. Johnson, “A Semantic Lexicon for Medical Language Processing”, JAMIA, 1999 May–Jun; 6(3): 205–218
7. Udo Hahn, “Tagging Medical Documents with High Accuracy”, Pacific Rim International Conference on Artificial Intelligence Auckland, Newzealand , pp. 852–861, 2004
8. Hans Paulussen, “DILEMMA-2: A Lemmatizer-Tagger for Medical Abstracts”, In Proceeings of ANLP, pp.141–146, 1992
9. Carol Friedman, “Automatic Structuring of Sublanguage Information,” London: IEA, 1986, pp. 85–102.
10. Emile C. Chi et al., “Processing Free-text Input to Obtain a Database of Medical Information”, In Proceedings of the 8th Annual ACM-SIGIR Conference, 1985
11. Udo Hahn, “Automatic Knowledge Acquisition from Medical Texts”, In Proceedings of the 1996 AMIA Annual Fall Symposium, pp.383–387, 1996
12. What is CDA?: http://www.h17.org.au/CDA.htm#CDA
13. David Elworthy, “Does Baum-Welch Re-estimation Help Taggers?”, Proceedings of the 27th ACL, 1989
14. Bernard Merialdo, “Tagging English Text with a Probabilistic Model”, Computational Linguistics 20.2, pp155–172, 1994
15. A. J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimal decoding algorithm”, IEEE Transactions of Information Theory 13, pp 260–269, 1967
16. Baum, L, “An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process”, Inequalities 3:1-8, 1972.