Integration of Low Level Linguistic Information for Clinical Document Semantic Tagging

Hyeju Jang*, Yun Jin**, and Sung Hyon Myaeng*
*Department of Computer Science, Information and Communications University, Daejeon, Korea
**Department of Computer Science, Chungnam National University, Daejeon, Korea
hjjang@icu.ac.kr, wkim@cnu.ac.kr, myaeng@icu.ac.kr

Abstract

We propose a semantic tagger that provides high level concept information for phrases based on several kinds of low level information about words in clinical narrative texts. The semantic tagging, based on Hidden Markov Model (HMM), is performed on the text that has been tagged with Unified Medical Language System (UMLS), Part-of-Speech (POS), and abbreviation tags. It reuses UMLS, POS, abbreviation, clue words, and numerical information to produce higher level concept information. Our unknown phrase guessing method for a robust tagger also uses the existing information calculated in the training data. In short, the semantic tagger gives more meaningful and abstract information by integrating different kinds of low-level information.

1. Introduction

Medical documents written by doctors are apt to have very different characteristics compared with other kinds of documents. A remarkable feature is that they contain many specialized technical terms in the domain. In addition, the documents have lots of abbreviations and non-alphanumeric symbols that can be understood only by those how have been trained in the small group.

Especially, the clinical documents written by Korean doctors have features distinguished from those written by the doctors using English as the mother tongue. The clinical documents written in Korea are written in both Korean and English. Usually, English is used for the medical terminologies, and Korean is used for some general nouns and most verbs. As a result, the grammars for English and Korean separately are not so much useful; they should be applied in a unique way to parse the text, thereby needing a way to integrate analysis results based on Korean and English grammars.

In this situation, it is hard to make full use of lexical and syntactic information of the Korean clinical texts when the purpose of the analysis is to be available to the users who can learn from patient records new knowledge without direct experiences. Even a syntactic parser can hardly be applied to such texts. The structural information in Korean clinical text that can be revealed by a grammar is not as important as word-level information. In other words, it may be sufficient to treat Korean clinical documents as a bag of words.

This paper describes a tagging system that uses the characteristics of words in order to yield high-level semantic tags for phrases in the clinical documents. The tags in this system are categories of information that phrases of medical records contain, such as symptom, therapy, and performance. To identify which category a phrase belongs to, the system uses the features of words in a phrase. It tracks down the relationships among a symptom, a therapy, and its performance, and retrieves past cases with which users want to know about a certain therapeutic method, for example.

The contribution of this research is that it provides a way of using existing information in clinical texts whose grammar is not necessarily based on either Korean or Korean grammar. The system annotates clinical text on phrases based on integration of the various existing knowledge embedded in the past patient records. It yields high-level semantic annotation on phrases with syntactic information and odd bits of semantic information. It also provides a method for guessing unknown phrases for more robust text processing, which in addition uses the information of the training data.

2. Related Works

The language used in a particular domain is called a sublanguage. The language dealt with in this research can also be regarded as a medical sublanguage since our research is restricted to the medical domain. There have been some papers which mentioned to the characteristics

0-7803-9788-6/06/$20.00 ©2006 IEEE.

of a medical sublanguage. Riochard Kittredge [1] talked about the factors of a sublanguage.

1) Restricted domain of reference
2) Restricted purpose and orientation
3) Restricted mode of communication
4)Community of participants sharing specialized knowledge

Emile C. Chi et all [2] and Sa Kwang Song [3] gave a talk about the characteristics of a medical sublanguage as well. Medical records have a lot of specialized medical words, abbreviations, and non-alphanumeric symbols which others but doctors can hardly understand the meaning of. For example, an upward arrow sometimes becomes 'increased'. In addition, "sl" is used as an abbreviation instead of "slight."

The popular and conventional approach of part-of-speech (POS) tagging systems is to use a Hidden Markov Model (HMM) [4] so as to find a most proper tag [5]. Some systems use a HMM with additional features. Julian Kupiec [6] and Dong Cutting et all [7] 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 equivalent classes.

Tagging systems in the medical field have focused on the lexical level of syntactic and semantic tagging. Patrick Ruch [9] and Stephen B. Johnson [9] performed semantic tagging on terms lexically using the Unified Medical Language System (UMLS). On the other hand, Udo Hahn et all [10] and Hans Paulussen [11] built POS taggers which categorized words syntactically.

There also have been the systems which extract information from the medical narratives [12, 13]. Friedman [12] defined six format types that characterize much of the information in the medical history sublanguage.

Sa Kwang Song [3] did research for abbreviation disambiguation in the medical documents, which is important since medical documents have a bunch of abbreviations.

3. Target Semantic Tag

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. In order to answer these questions, we use the "narrative" data part of CDA (Clinical Data Architecture) documents as the knowledge source, which came from SNHU, since they contain the past treatment history of individual patients. 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.

4. System Architecture

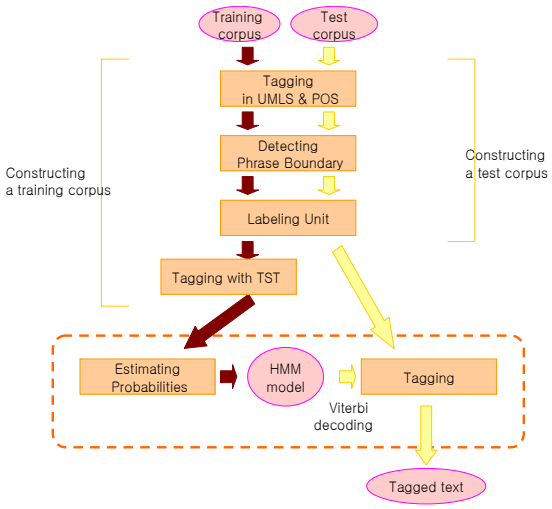


Fig. 1. The system architecture for the TST tagger

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 HMM. HMM models in other applications like POS tagging have used the grammar rules or syntactic patterns for state transitions and emissions to find the most probable sequences. 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 architecture of the system using HMM is shown in Fig. 1. The whole architecture is largely divided into the training stage and the tagging stage.

4.1. Training Stage

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 [3] in the same project. We treat abbreviations in a special way because they are sometimes ambiguous and not handled properly by either UMLS or POS taggers.

After tagging to prepare the ground, the text is divided as a unit of a phrase. 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. So, 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.

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.

After connecting with equivalence classes, the tagging is done manually for the training corpus. The training corpus is completed with manual tagging finally.

Then, frequencies of words/phrases/tags are counted to estimate the probabilities required for a HMM model so as to use tagged training data. 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 this method because its accuracy is much higher than that of the Baum-Welch method [15, 16, 17].

TABLE I
. Equivalence classes on 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)
Numeric for Date	Date of the event, time-order information
Numeric for prescription	The frequency of taking medication, does information
unknown	neither clue word nor UMLS tag

4.2. Tagging Stage

It is performed on the test corpus in the same way as in the training stage to do POS, UMLS, and abbreviation tagging , to divide as phrases, and to connect phrases with the equivalence classes. And then, the system finds a most probable tag sequence using the Viterbi algorithm [18] using the HMM model constructed in the training stage.

5. Equivalence Classes

Equivalence classes are the key part because the observance of HMM in this system is composed of a combination of equivalence classes. In addition, equivalence classes are critical in that they reuse the basic tagging information and other resources included in the text. A phrase is transformed into the combination of equivalence classes. Fig.2. shows how it is transformed with the example. Since equivalence classes represent the information of a phrase, they say how well it can reflect the characteristics of data. Accordingly, it can determine the performance of tagging directly.

This research uses the information of UMLS, clue words, abbreviations, and numeric data for equivalence

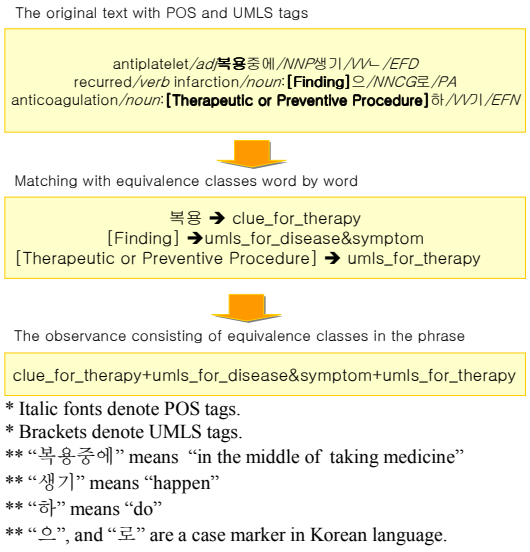
classes. They are components which can say what kind of meaning the phrase has.

For a number of medical terms, UMLS tagging is used to annotate what kind of medical terms they are. The performance of UMLS tagging are 42.95% for thepapeutic or preventive procedure which mostly represents 'Therapy', and 77.68% for disease or syndrome which is for 'Syndrome'.

The system classifies UMLS by the meaning for the purpose of tagging. UMLS tags are classified to cause, disease or symptom, and therapy because this system targets three tags of 'Symptom', 'Therapy', and 'Performance'. In the mean time, other UMLS tags are ignored since they are not related with the TST.

Aside from UMLS tags, clinical narrative data consists of many abbreviations and numeric data. It needs to be processed for each part separately since they are ambiguous in many cases.

Abbreviations in the corpus are processed based on the research [3] in the same project. Disambiguation of abbreviations is handled with the help of semantic abstraction of symbols and numeric terms.



* 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

Numeric data in CDA documents are important because of amounts and meaning. CDA documents in this research include around 6.03% of numeric resources. These numerical symbols are used for date or numbering, and prescription related numbers largely. For date or numbering, there are dates of examination, operation, and ambulatory care visit. For prescription, numbers are used as orders, medical dosage, the number of times of taking medicine, and so on.

The process of numeric data takes advantage of the pattern rules which was made by analyzing the actual usage pattern according to the above two meaningful classification. For example, in the case of date, there are patterns like "1998-3-2", "98.3.2", "98/3/2", and "3.2" and in the case of numbering and prescription, there were patterns like "#1, #2, ,,,", and "30 mg". There were difficult numbers to be processed. For instance, "3/2" or "3.2" can be a date or something else. In this research, resolving the ambiguity of numeric is not an issue since it is just one resource of many resources. The performance of numeric tagging is 80.5% for dates relating to numbers and 86% for numbering and prescription relating to numbers.

6. Guessing Unknown Phrases

For guessing unknown phrases, we also use the existing information for the purpose of attaching an appropriate tag to the phrase.

Unknown 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 labels. 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 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.

7. Performance

7.1. 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 [14].

We picked 300 narrative 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.

7.2. Experiments and Results

The level of accuracy of our system is calculated as the number of correct tags per the total number of tags.

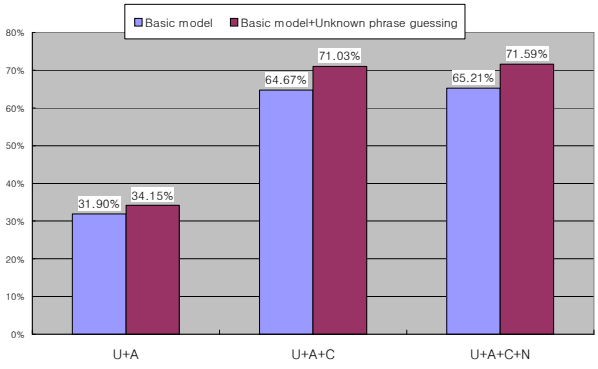


Fig. 3. Performance comparison

We compared several cases of the performances of each system using different equivalence classes. Fig. 3. show the comparison of the performance for each case of UMLS+Abbreviation data, UMLS+Abbreviation+Clue words data, and UMLS+Abbreviation+Clue words+Numeric data. We can see the performance increase with the more features for equivalence classes. Even with integrating inaccurate data, the performance of the whole system is getting better slightly.

7.2.1. UMLS + Abbreviation. The baseline system uses the combination of UMLS tags of words composing a phrase. This experiment shows how the accuracy is when the tagging only depends on UMLS and abbreviation tags.

7.2.2. UMLS + Abbreviation + Clue Words. This experiment shows how the accuracy increases when we add the equivalence class concept using UMLS, abbreviation and clue words to represent a phrase. It gives almost two times improvement compared with the baseline case.

7.2.3. UMLS + Abbreviation + Clue Words + Numeric terms. This experiment shows the importance of numeric terms which occupy about 6.03% in documents. It represents slight improvement compared with U+A+C case. We guess the reason may be the categories for numeric data are not enough.

8. Conclusion

We showed our method of building a semantic tagger for medical documents using different kinds of information with HMM. It was shown that different kinds of information can be used as the observances in HMM. Even the unknown phrase guessing problem can be solved with the same mechanism by using such information corresponding to the component words. For future work, we plan to expand the research by using other machine learning methods such as Conditional Random Field as a way of improving the performance.

9. Acknowledgement

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

10. References

[1] Richard Kittredge, "5. Sublanguage", Americal Journal of Computational Linguistics, 1982

[2] Emile C. Chi et all, "Processing Free-text Input to Obtain a Database of Medical Information", In Proceedings of the 8th Annual ACM-SIGIR Conference, 1985

[3] Sa Kwang, Song, "Abbreviation Disambiguation Using Semantic Abstraction of Symbols and Numeric Terms,", 2005, IEEE NLP-KE

[4] L.R.Rabiner et all, "An Introduction to Hidden Markov Models", IEEE ASSP Magazine, 1986

[5] Linda Van Guilder, "Automated Part of Speech Tagging:A Brief Overview", Handout for LING361, 1995

[6] Julian Kupiec, "Robust part-of-speech tagging using a hidden Markov model", Computer Speech and Language, pp. 225–242, 1992.

[7] Doug Cutting et all, "A Practical Part-of-Speech Tagger", In Proceedings of the 3rd ACL, pp.133–140, 1992
[8] Patrick Ruch, "MEDTAG: Tag-like Semantics for Medical Document Indexing", In Proceedings of AMIA'99, pp.35–42
[9] Stephen B. Johnson, "A Semantic Lexicon for Medical Language Processing", J Am Med Inform Assoc. 1999 May–Jun; 6(3): 205–218
[10] Udo Hahn, "Tagging Medical Documents with High Accuracy", Pacific Rim International Conference on Artificial Intelligence Auckland, Newzealand , pp. 852–861, 2004
[11] Hans Paulussen, "DILEMMA-2: A Lemmatizer-Tagger for Medical Abstracts", In Proceeings of ANLP, pp.141–146, 1992
[12] Carol Friedman, "Automatic Structuring of Sublanguage Information," London: IEA, 1986, pp. 85–102.
[13] Udo Hahn, "Automatic Knowledge Acquisition from Medical Texts", In Proceedings of the 1996 AMIA Annual Fall Symposium, pp.383–387, 1996
[14] What is CDA?: http://www.h17.org.au/CDA.htm#CDA
[15] Baum, L, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process", Inequalities 3:1-8, 1972.
[16] David Elworthy, "Does Baum-Welch Re-estimation Help Taggers?", Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 1989
[17] Bernard Merialdo, "Tagging English Text with a Probabilistic Model", Computational Linguistics 20.2, pp155–172, 1994
[18] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm", IEEE Transactions of Information Theory 13, pp 260–269, 1967