Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Semantic Tagging for Medical Knowledge Tracking

Hyeju Jang*, Sa Kwang Song**, and Sung Hyon Myaeng*
*Department of Computer Science, Information and Communications University, Daejeon, Korea
**Electronics and Telecommunications Research Institute, Daejeon, Korea

***Abstract*—**We propose a semantic tagger that provides high level concept information for phrases in clinical documents, which enriches medical information tracking system that support decision making or quality assurance of medical treatment. In this paper, we have tried to deal with patient records written by doctors rather than well-formed documents such as Medline abstracts. In addition, annotating clinical text on phrases semantically rather than syntactically has been attempted, which are at higher level granularity than words that have been the target for most tagging work.

I. Introduction

Many researches in biomedical domain as well as computer science have worked for mining useful knowledge from well-formed documents such as Medline abstract or thesis paper. Document analysis and knowledge acquisition with well-formed documents is relatively convenient and better performed than with coarse-grained documents because it enables applying deep natural language processing techniques into document.

Especially in medical domain, patient records written by doctors have very different and intractable characteristics in that they usually contains large amount of numeric terms and symbols as well as linguistically ill-structured sentences. That is the reason we can hardly find related works on coarse-grained patient records, even though they 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 [13].

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.

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 [2] for semantic annotation.

II. 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 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 [11] and Sa Kwang Song [14] 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 HMM model so as to find a most proper tag [3]. Some systems use a HMM with additional features. Julian Kupiec [4] and Dong Cutting et all [5] 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 [6] and Stephen B. Johnson [7] performed semantic tagging on terms lexically using the Unified Medical Language System (UMLS). On the other hand, Udo Hahn et all [8] and Hans Paulussen [9] built POS taggers which categorized words syntactically.

There also have been the systems which extract information from the medical narratives [10, 11, 12]. Friedman [10, 11] defined six format types that characterize much of the information in the medical history sublanguage.

Sa Kwang Song [14] did research for abbreviation disambiguation in the medical documents, which is important since medical documents have a bunch of abbreviations.

III. SCENARIO

The following is a scenario to show problem cases and expected solutions with the proposed system, CDA tracker.

If a doctor wants to know "What are the performance characteristics of syrup in the setting of Malignant Pleural Effusion?", he/she can use a knowledge tracking system, CDA tracker, whose goal is to provide useful information to doctors based on document understanding process on clinical documents. It means that the tracker retrieves the CDA documents in which the relevant parts are highlighted and post-processed by real-time clustering method.

To bring answers to the user, the CDA tracker consists of several parts, which are shown in Fig. 1. Semantic phrase tagging is one of the parts of the whole system and it is processed in the preprocessing stage in the architecture. The tracker does its job based on the documents which is finished with semantic phrase tagging as well as other primitive tagging.

Semantic phrase tagging in here is in order to track an answer to the user query. It is differentiated from anything people have ever done in the medical field such as UMLS semantic tagging [6] and information extraction [10]. Tracking is to retrieve answers to a specific problem rather than analyzing the text in a general usage. So, it can be said that the problem here is different from other research issues which have been dealt with in the medical information area. For this reason, there is no other approach in the same problem, or tracking yet.

And, because semantic tagging here is for tracking, it assigns concepts like "Therapy", "Symptom", and "Performance" to the phrases, which are the concepts in a certain user query targeted to the tracker. It is direct to give answers to a user because its tags reflect a user's need with the similar level concept for the CDA tracker. It is based on other primitive tagging such as POS and UMLS tagging which simply show syntactic and semantic categories of a word, respectively. For example, the question mentioned above, "What are the performance characteristics of syrup in the setting of Malignant Pleural Effusion?", "syrup" is a therapy, "Malignant Pleural Effusion" is a symptom, and the user wants to know the performance of syrup on malignant pleural effusion. Semantic tagging assigns those essential tags to the documents.

Fig. 2 shows an input text and the result of semantic tagging. The text is divided into phrases. And then, they are tagged with corresponding semantic categories.

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). 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.

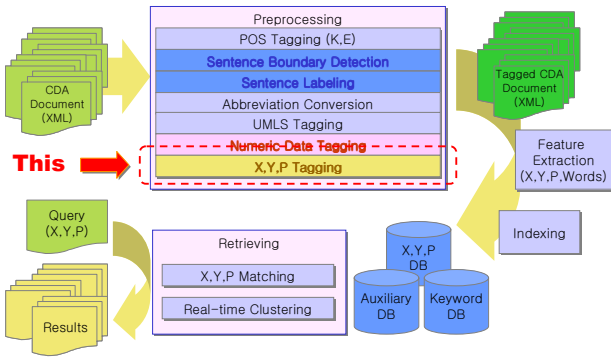

Fig. 1. CDA tracker architecture

Ativan therapy하며 방에 재움. 점차 Liver failure 진행되어 사망

<Therapy><u>**Ativan**/noun:[Organic Chemical]
therapy/noun:[Therapeutic or Preventive Procedure]**하**/VV**며**/EFC</u></Therapy>
<Therapy><u>**방**/NNCG**에**/PA**재우**/VV**ㅁ**/EFN</u></Therapy>
<Performance><u>**점차**/AA**Liver**/noun:[Body Part, Organ, or Organ Component]
failure/noun:[Functional Concept]**진행**/NNCV</u></Performance>
<Garbage><u>**되**/XSVV**어**/EFC</u></Garbage>
<Performance><u>**사망**/NNCV</u></Performance>

* <u> .. </u> is tags for representing a phrase
* <Therapy>, <Performance> and <Garbage> are target semantic tags.

Fig. 2. An example of semantic phrase tagging

IV. METHODOLOGY

A. Training Stage

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.

2) Tagging for Abbreviations: Abbreviations in the corpus are processed based on the research [14] 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. Numeric terms are also handled here.

3) Detecting Phrase Boundary: 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.

4) 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.

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

6) Estimating Probabilities: There are two training methods used in current tagging systems. One is to use the Baum-Welch algorithm [18] 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 [15] and Bernard Merialdo [16] 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.

B. Tagging Stage

1) Tagging in UMLS & POS: It is performed on the test corpus in the same way as in the training stage.

2) Tagging for Abbreviations: Abbreviations in the test corpus is processed in the same way as in the training stage.

3) Detecting Phrase Boundary: This process is also done in the same way as in the training stage.

4) Labeling Unit with Equivalence Class: Phrase units are tagged with the equivalence class labels in the same way as in the training stage.

5) Tagging: The system finds a most probable tag sequence using the Viterbi algorithm [17] using the HMM model constructed in the training stage.

6) Tagging of unknown phrase units: 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.

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)
unknown	neither clue word nor UMLS tag

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.

V. EXPERIMENTS AND RESULTS

A. 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 [13]. Fig. 3 is a snippet of a CDA document.

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. Because the amount of corpus is small, we used 3-fold validation.

<Select_증요필드><doc_id>00525</doc_id><주호소C_C>RVOTO에 대한 Mx.</주호소C_C><입원시경과>유지</입원시경과><진단명> 9024 Tetralogy Of Fallot , 6459 Seizure ,</진단명><문제목록> TOF , HIE ,</문제목록>

<증요검사소견>Echocardiography :: [증간 보고 (검사일자 : 2002-11-23)] RVOT velocity : 1m/sec. PI (+) : trivial TR (-). MRI :: [최종 보고 (판독일자 : 2002-10-18)] 2002-10-07 BRAIN MRI (ROUTINE) [Finding] Diffuse 한 cortical gyral swelling 소견이 있음. Thalamus 와 lentiform nucleus 의 signal intensity 가 T2WI 에서 증가되어 있음. Hypoxic ischemic encephalopathy 에 consistent 한 finding 임. Enhance 를 시행하였을 때 gyral enhancement 의 소견이 있음. Parenchyma 에 focal lesion 은 보이지 않음. [최종 보고 (판독일자 : 2002-11-23)] 2002-11-15 BRAIN MRI (ROUTINE) [Finding] 2002-10-07 brain MRI 와 비교판독하겠음. 이전 양측 cerebral hemisphere 에 특히 right side 에 좀더 predominant 한 양상의 T2 high signal intensity lesion 은 FLAIR image 상에서 gyrus 를 따라서 특히 high convexity area 및 parasagittal area 에 일부에도 linear 한 high signal intensity 를 보이고 enhanced 한 scan 에서 diffuse 하게 gyrus 를 따라서 enhancement 가 되고 있음. 이는 이전의 hypoxic ischemic encephalopathy 에 nature course 로 laminanecrosis 에 의한 signal intensity change 로 생각되며 새로 생긴 focal lesion 은 없음. 양측 brain parenchyma 에 mild 한 atrophic change 가 동반되어 있음. 판독의사 : 손규리/김인원 [Diagnosis] Diffuse gyral enhancement along both cerebral hemisphere, predominant in the right side. --> No definite evidence of newly developed brain parenchymal lesion. Natural course of hypoxic ischemic encephalopathy, most likely. EEG findings ------중략------</증요검사소견> <퇴원형태>귀가</퇴원형태>

<퇴원시처방>투약 Phenobarbital 30mg tab하나 (Phenobarbital) 75 mg bid pc 1 일 2회 * 30일 Pheton susp 25mg/ml(Diphenylhydantoin) 75 mg bid pc 1 일 2회 * 30일 Ebiose powder (Ebiose) 0.25 g tid pc 1 일 3회 * 30 일</퇴원시처방> <외래예약정보>1개월 후 소아과 채종희님 외래 f/u</외래예약정보>

<현병력>TOF 진단받고 91년 Op 받은 후 외래 (소아 TS) F/U 받던 환아로 2002/10/1 ReOp 시행 (d/t RVOTO) 하였다. PICU에서 care 받던 중 PostOp 6시간 째 Seizure 발생하였다. seizure는 GTC type으로 5분간, Lt. arm 흔드는 양상으로 5분간하였다. Pb, DPH loading 하고 midazolam continuous infusion 한 후 seizure control 되었고, 이후 maintenance로 유지하며 seizure free 상태이나 환아 말 못하는 상태로 bed ridden 하고 있던 중 2002/10/13 일반병동으로 전동된 후 10/14 소아과 천동 되었다. 10/2 시행한 portable EEG상 medium amplitude theta and delta activities with occasional atypical sharp wave from the Rt. and lt. frontal lobe 에 보였고, 10/7 시행한 brain MRI 상 diffuse gyral swelling & thalamus, lentiform nucleus 에 high signal 보여 hypoxic ischemic encephalopathy에 부합된다 하였다.</현병력>

<신체검진>2002/12/28 on duty 당시의 검진 결과임 G/A : Not so ill looking V/S : 103/51mmHg- 105/min- 24/min- 36.5도 HEENT: NCP not anemic, anicteric, conjunctival injection (-/-) TM : both not hyperemic oral cavity내는 환아 협조가 되지 않아 평가하지 못함 Neck: L/N(-/-) Chest: symm. expansion without retraction OP scar(+) RHB without murmur CBS without rale, wheezing Abd: soft, flat Normoactive Bowel sound T/RT(-/-) L/S(-/-) B&Ext: P/C/C(-/-), no CVA tenderness Skin : No petechiae, No purpura Anogenitalia : No gross anomaly Perianal abscess (-) N/Exam> MSE alert, 검사자에 대해 친근한 태도 보임 질문에 대해 때로는 적합하고, 때로는 적합하지 않은 응답을 함 ------중략------ </신체검진>

<입원후경과_기타정보>02/10/1 PVR 시행함, PICU 전동됨 02/10/2 Seizure attack 나타남 AED 투여 및 EEG w/u 시행함 02/10/6 Midazolam C.I --> OFF 02/10/7 Brain MRI 시행함 02/10/13 ward로 Tf. 02/10/14 소아과로 Tf. 02/10/15 RM consultation :: tilting table & PROM exercise 02/10/16 Antibiotics D/C 02/10/17 영양과 consultation :: 1600kcal까지 증량하자 02/10/22 Fever로 cefotaxime start ------중략------</입원후경과_기타정보></Select_증요필드>

Fig.3. A snippet of a CDA document

B. *Result*

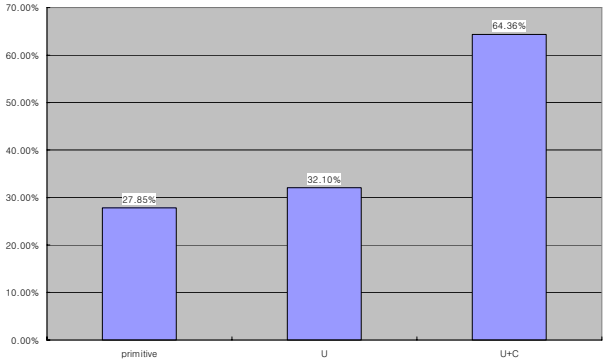

Fig. 4. The results of a basic HMM model and a HMM model with equivalence classes

The level of accuracy of our system is calculated as the number of correct tags per the total number of tags. Fig.4 shows the comparison of the basic HMM model with the one adopting the idea of this paper. Although the result of the system is not as good as expected, it is promising with further improvement since it already shows the HMM model we designed is much more efficient than the HMM model using primitive tags.

There is a serious limitation in building a training corpus manually. Most of all, the amount of the training corpus is not enough. To show the method proposed in this research is efficient, a certain level amount of the training corpus is required. Unfortunately, lack of the medical experts who can build training corpus manually made this limitation. When we can use a enough amount of training corpus by domain experts, we expect that the experiment result can show the efficiency of our method better.

VI. 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.

ACKNOWLEDGMENT

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

REFERENCES

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

[2] L.R.Rabiner et all, “An Introduction to Hidden Markov Models”, IEEE ASSP Magazine, 1986

[3] Linda Van Guilder, “Automated Part of Speech Tagging:A Brief Overview”, Handout for LING361, 1995

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

[5] Doug Cutting et all, “A Practical Part-of-Speech Tagger”, In Proceedings of the 3rd ACL, pp.133–140, 1992

[6] Patrick Ruch, “MEDTAG: Tag-like Semantics for Medical Document Indexing”, In Proceedings of AMIA’99, pp.35–42

[7] Stephen B. Johnson, “A Semantic Lexicon for Medical Language Processing”, J Am Med Inform Assoc. 1999 May–Jun; 6(3): 205–218

[8] Udo Hahn, “Tagging Medical Documents with High Accuracy”, Pacific Rim International Conference on Artificial Intelligence Auckland, Newzealand , pp. 852− 861, 2004

[9] Hans Paulussen, “DILEMMA-2: A Lemmatizer-Tagger for Medical Abstracts”, In Proceeings of ANLP, pp.141− 146, 1992

[10] Carol Friedman, “Automatic Structuring of Sublanguage Information,” London: IEA, 1986, pp. 85–102.

[11] 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

[12] Udo Hahn, “Automatic Knowledge Acquisition from Medical Texts”, In Proceedings of the 1996 AMIA Annual Fall Symposium, pp.383− 387, 1996

[13] What is CDA?: http://www.h17.org.au/CDA.htm#CDA

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

[15] David Elworthy, “Does Baum-Welch Re-estimation Help Taggers?”, Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 1989

[16] Bernard Merialdo, “Tagging English Text with a Probabilistic Model”, Computational Linguistics 20.2, pp155–172, 1994

[17] A. J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimal decoding algorithm”, IEEE Transactions of Information Theory 13, pp 260–269, 1967

[18] Baum, L, “An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process”, Inequalities 3:1-8, 1972.