Original Article

Healthc Inform Res. 2011 September;17(3):150-155.
http://dx.doi.org/10.4258/hir.2011.17.3.150
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Recognizing Temporal Information in Korean Clinical Narratives through Text Normalization

Youngho Kim, MS[1], Jinwook Choi, MD, PhD[2]

[1]Interdesciplinary Program of Bioengineering, College of Engineering; [2]Department of Biomedical Engineering, College of Medicine, Seoul National University, Seoul, Korea

Objectives: Acquiring temporal information is important because knowledge in clinical narratives is time-sensitive. In this paper, we describe an approach that can be used to extract the temporal information found in Korean clinical narrative texts. **Methods:** We developed a two-stage system, which employs an exhaustive text analysis phase and a temporal expression recognition phase. Since our target document may include tokens that are made up of both Korean and English text joined together, the minimal semantic units are analyzed and then separated from the concatenated phrases and linguistic derivations within a token using a corpus-based approach to decompose complex tokens. A finite state machine is then used on the minimal semantic units in order to find phrases that possess time-related information. **Results:** In the experiment, the temporal expressions within Korean clinical narratives were extracted using our system. The system performance was evaluated through the use of 100 discharge summaries from Seoul National University Hospital containing a total of 805 temporal expressions. Our system scored a phrase-level precision and recall of 0.895 and 0.919, respectively. **Conclusions:** Finding information in Korean clinical narrative is challenging task, since the text is written in both Korean and English and frequently omits syntactic elements and word spacing, which makes it extremely noisy. This study presents an effective method that can be used to aquire the temporal information found in Korean clinical documents.

Keywords: Medical Informatics, Information Processing, Multilingualism, Medical Record, Automated Pattern Recognition

I. Introduction

A large amount of textual data has been increasingly available along the use of modern electronic health record system. While structured data such as lab test results or coded chief complaints provide direct information in clinical research, unstructured free text such as discharge summaries were hardly used because much of human efforts were needed to deliver information from the text. To support these tasks, information extraction systems in the clinical field have been profoundly researched.

Among many information of interest, temporal information has been an important component in the clinical information applications. Clinical unstructured text, such as progress notes, discharge summaries contain important clinical

Submitted: September 7, 2011
Revised: September 14, 2011
Accepted: September 20, 2011

Corresponding Author
Jinwook Choi, MD, PhD
Department of Biomedical Engineering, College of Medicine, Seoul National University, 103 Daehak-ro, Jongno-gu, Seoul 110-744, Korea. Tel: +82-2-2072-3421, Fax: +82-2-745-7870, E-mail: jinchoi@snu.ac.kr

events. Every clinical event, such as onset of a symptom, order of medications, diagnosis and treatments has its specific point of time, and temporal expressions that describe those points of time are located near the event description. These temporal expressions of events can enhance both the content selection and the summary generation process. As a result, extracting temporal information has been part of many important applications such as clinical information retrieval of certain time period [1,2], summarizing chronological events [3,4] and providing information for time-dependent decision support system [5].

A number of approaches and systems have been developed to extract temporal information from natural language texts. Most of their work aimed to extract temporal information from the newswire text. Verhagen et al. [6] developed a system for extracting temporal information from newswire text. Their system annotated absolute, relative, and durational temporal expressions and modifiers. They defined an absolute time as a temporal expression that does not dependent on spoken or written context, and a relative time as one that dependent on a referenced time point. They used natural language techniques such as part-of-speech tagging, noun phrase chunking to process English text. The performance of their system has been evaluated, and they scored F-measure of 0.85. Another approach was done by Ahn et al. which applied machine-learning approach for the recognition and normalization task [7]. They first recognized temporal information from the text with machine-learned classifier then performed normalization task that transforms temporal information to timestamps. They evaluated their system with several combination of rule-based and machine-learning based approach, and they concluded that the combination of those methods can outperform approaches that are strictly rule-based.

In clinical field, researches on extracting temporal information focused to discharge summaries. Zhou et al. [8] published several studies that dealt with temporal information in medical text, which includes discussions about annotation scheme, system architecture and evaluation of their temporal information extraction system. They evaluated the performance of their system and they reported the accuracy of 95%. However the result was incomparable to former systems due to the different metrics they used. Notably, they listed several challenges in extracting temporal information in medical text, which includes diverse temporal expressions, medical grammar, pragmatics and ubiquitous ambiguities. Their study shows the domain-specific problem in temporal analysis task, meaning that approaches with standard natural language processing techniques that used to be applied in the newswire text do not comply well in the clinical domain. Our target document is Korean discharge summaries, and its text contains both Korean and English terms. Most of key medical terms such as disease name, symptom and anatomical locations are often written in English, and Korean language tends to be used for describing patients' complaints and life behaviors. In our target document, time information was described in mixture of both languages, which made the task more challenging.

Clinical documents written in English are syntactically well-structured, and this grammatical soundness allows applying standard natural language processing (NLP) approaches such as white-space tokenization, sentence boundary detection and part-of-speech tagging in the process. Syntactic structure in Korean discharge summary is much less distinctive, since the document is mostly summarization of patient's history. Documents are written with no sentence break, almost grammar-free and omitted word spacing, mainly due to the timely nature of author's situation. Figure 1A shows a fragment of text from discharge summary as an example. In this context, analyzing syntactic structure is not feasible. Therefore in our approach, text processing was mainly based on lexical feature and corpus-based heuristics which uses large lookup lists and corpus statistics.

II. Methods

Our system architecture for extracting temporal expression is presented in Figure 2. The whole process consists of mainly two components, which are token analyzer and temporal expression recognizer. The former component of the system deals with the text tokenization. The tokenization is the beginning of all NLP process. A word is the basic

Figure 1. (A) An example of temporal expression marked from the medical history of Korean discharge summary. (B) Same text translated into English. Note that the arrangement of word is changed because the word order of Korean is different from English.

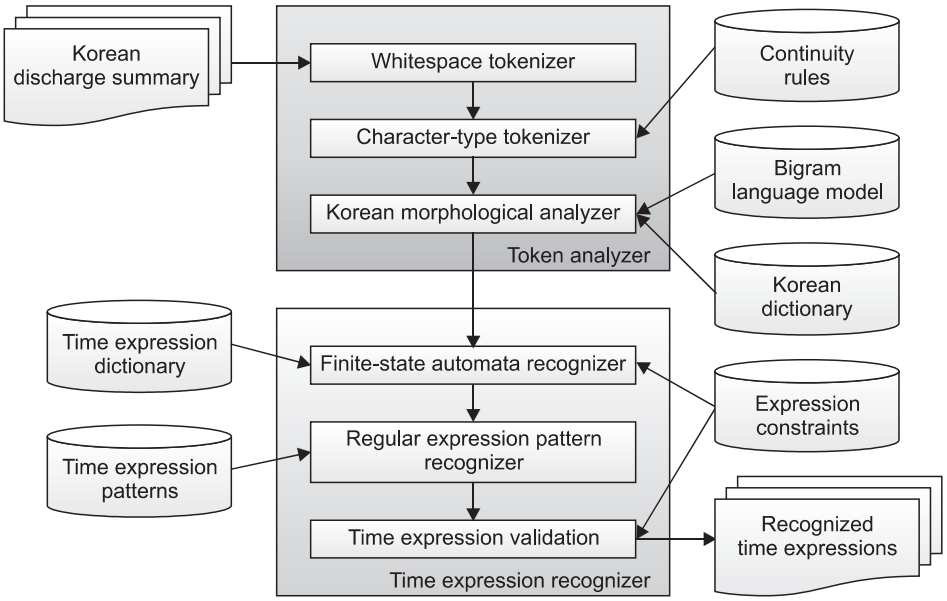

Figure 2. The system architecture for temporal expression extraction in Korean clinical narrative.

component of the English literature, and normally words are clearly separated with whitespaces and punctuations in the text. Therefore, separation between those characters is considered to be sufficient in the most of text processing systems. However, applying same method to Korean text is not appropriate for several reasons. First of all, Korean language makes use of various affixes, pre- and postpositions that are bound morphemes that combine with word stem. The components such as prepositions are written in separate word in English text, so white space tokenizing could be effective in the processing. There are no such markers that divide affixes from the stem in Korean. To resolve the tokenizing problem, exhaustive token analyzer was designed for our system.

Also the word spacing may be omitted without disturbing the reader's understanding, though being incorrect to the grammar. A token such as "내원수개월전에 (before several months to the visit)" is actually contains 5 word elements joined together. Though it should be processed into 5 parts according to the grammar, Korean readers easily recognize the meaning without ambiguity. These derivation and omission are more prevalent in the summary text, because the authors of the text tend to shorten the text.

At the glance of tokens, character type was most obvious feature for dividing a token. Four character types were set, which were English and Korean letters, numbers and symbols. In the beginning of the process, a token was decomposed into characters. Characters of same type were joined while characters between different classes were divided. During this step, numeric expressions were unintentionally divided into a couple of smaller parts, for example, "3.4→3/./4", so those tokens were restored again after separation. We implemented continuity rules to restore those pieces into original expression.

As the next procedure to the character-type tokenizer, morphological analyzer that based on knowledge-based approach combined with corpus statistics was implemented. Initially, the longest word matching algorithm with Korean dictionary was used to separate components from a Korean token, though it was turned out to produce many false matches in the result. For example, a token "수차례 (several times)" which omits spacing between "수 (several) 차례 (times)" had to be separated as it is. However, when the algorithm falsely matched longer morpheme in the beginning "수차 (turbine)", the remaining "례 (etiquette)" also resulted as a false match. This shows an example of error propagation, and we tried to minimize the problem through using corpus statistics over dictionary lookup results to determine most probable separation of words. In the process, bigrams and its corpus probabilities were derived from Sejong Corpus, which is grammatically supervised national linguistic corpus [9]. Every text in the corpus was grammatically corrected by linguistic experts, as well as word spacing. We used the SRILM (SRI Language Model Tookit) to build the bigram language model from the corpus [10].

The process for estimating correctly separated case began with generating all case of combinations using each character in a token, and 2^n-1 cases were generated for each n-character token. Algorithm then matched each separated token against dictionary. All bigrams from each separation were also matched against language model to retrieve probability information. Based on the information, certainty scores for each separation were calculated. The scoring function for

http://dx.doi.org/10.4258/hir.2011.17.3.150

this step is calculated as follows:

$$\text{Certainty Score}(S) = \frac{1}{|S|}\sum_{w\in S}\alpha_{|w|}I_d(w) \times \frac{1}{|S|-1}\sum_{w,w_p\in S}\beta_{|w|}I_{bi}(w,w_p) \times \gamma\sum_{w,w_p\in S}p_{bi}(w,w_p)$$

where S is set of tokens from each separated case. |S| is the size of set S, w is a tokens from set S while wp is preceding token of w, Id(w) is an identity function that results 1 when token w is found in a target lookup list d, which is unigram dictionary. Among separated cases, the highest scoring one is chosen as a correct case.

In the initial result, this algorithm tend to separate all the recognizable word into tokens, for example, "종합병원(general hospital)" into "종합(comprehensive)" and "병원(hospital)" because shorter words are tend to be used more frequently. To prevent over-separation and give tendency to produce longer tokens, parameter α, β, γ were empirically adjusted to give precedence to the longest words. Since this algorithm calculates certainty of all possible cases to perform correct tokenization, theoretically it produces optimal result. Unlike the longest sequence matching method, this algorithm showed robustness against the out of vocabulary cases by exploiting information of surrounding tokens.

Through our normalization of the token, we could reduce significant number of pattern variations that could be resulted from composition and derivation of Korean grammar. We designed a rule-based classifier that based on finite state automata (FSA) and regular expressions over temporal expression of the normalized text.

FSA recognizer operates on the sequence of inputs. Temporal expressions also can be viewed as a sequence of specific set of words. For an example, a sequence of tokens "3/개월/이전/부터(since 3 months before)" contains a digit "3", a time-related word "개월(months)" followed by an adverb "이전(before)" and postposition "부터(since)". These tokens can be easily recognized with matching against a lexicon. The lexicon was split into 4 category based on the part-of-speech and order of the words. For each state, only a specific lexicon in a category was tested for an acceptance. The recognition of a phrase began with the lexicon for the start state, and when the phrase satisfies the conditions of subsequent states, the phrase was recognized as a temporal expression. This simple FSA covered various temporal expressions, from short ones such as "'/90 (the year of 1990)", "1/YA (1 year ago)" to extended phrases such as "2/Y/3/MA (2 years and 3 months ago)", "약/2/-/3/개월/전/부터(roughly since 2-3 months ago)".

87.12 Ccr = 176 mL/min c-peptide 29.9 HbA1c 11.5% 88
left flank pain develop, 소주 1.5병 마신 후 hematemesis
있어 본원 ER visit

Figure 3. Temporal expressions along with the lab test results in the summary text.

Under the support of token analyzer, the recognition was performed well without proper spacing of the text.

There were exceptions that could not be identified by FSA. These phrases were mostly one word expression, which cannot form a sequence. Shorthanded date expressions with digits, such as "90.8", "2/1990" were the majority of the cases. One word expressions in Korean language such as "최근(recently)", "금일(today)" were also another majority cases. During recognition, 20 regular expressions and a lexicon for the one-word expressions were compiled to identify those expressions.

In the text, there were expressions such as "87.12" or "88" since authors of the summary frequently omit date-related words to shorten the writing. For example, the year before 2000 were frequently described omitting first two digits throughout our collection. Those short-handed expressions were just numbers in most cases, therefore may be interpreted as both measurement of the test result and partial temporal expression of a certain time point. As can be observed in Figure 3, the discharge summary contained a lot of test results in the text to describe patients' status, and the recognition may result a lot of possible false matches without proper post-processing. Those measurements are usually followed by a unit, or specified next to the name of tests. In order to prevent recognition of those measurements as a temporal expression, we implemented a rule-based validator to exclude those expressions based on the accompanying words in a window. Those test item patterns and units were compiled from our corpus as well as public source on the web.

III. Results

With the permission of the Seoul National University Hospital, we obtained a collection of discharge summaries from the EMR system among the patient records visited during 2003. Non-textual sections were removed from the documents, and the de-identified discharge summary only contained the date of the patient visit and text description of patient's past medical history. We manually annotated time information in the randomly selected 100 discharge summaries. The temporal expression included absolute and relative

date, and included postpositions that have specific meaning (e.g., -까지 "until", -부터 "since") and affixes (e.g., 약 "about"). Total 805 temporal expressions were annotated as a result. The result of the system is compared against manually annotated gold standards. The portion of the instances that correctly recognized by the system is interpreted as a precision, and the portion of correctly recognized instances out of the gold standard is interpreted as a recall. However, differ from the standard retrieval task, proper recognition of temporal expressions requires another aspect to be considered - the extent of the recognition. For this aspect, the measure of phrase-level precision (PP) and phrase-level recall (PR) was adapted from i2b2 NLP competition [11].

In the evaluation, the system found total 827 temporal expressions in the test set of 100 discharge summaries. Among those, 740 temporal expressions were correctly matched to the gold standard set, total 805 expressions. For each document, PP and PR were evaluated based on the result. Macro-averaged measures were calculated through averaging PP and PR over documents. Micro-averaged result was simply calculated from the sum of the results. The values are summarized in the Table 1.

We did another experiment to see the validation of the temporal expression could make differences in the system performance. Generally, temporal extraction systems did not consider this validation procedure, because their corpus rarely shows similar pattern in the text [6,7,12]. In the clinical setting, lab test measurements and dosage expressions were prevalent in the text, thus validations on those expression could reduce ambiguity in determining the meaning of the expressions. Comparison results of validations are summarized in the Table 2, and we could see the validation of the temporal expressions improved the precision of the system. According to the experiment, we confirmed that the simple heuristics that considers corpus characteristics can improve the overall system performance. Systems that deal with specific domain texts could be less effective when this kind of difference was not considered in the process.

IV. Discussion

With the use of the validator, we could filter out measurements for the lab tests, and false-positives were decreased with the increasing number of the constraints. With this method, we could exclude a lot of false-positives, and improved the precision, and achieved the F-score of 0.907 in micro-averaged results. Although our target documents were far noisier than general newswire articles, our system showed similar level of performance. Verhagen et al. [6] evaluated their system, and reported the recognition performance of 0.85 as F-score. Ahn et al. [7] also reported their recognition performance which was F-score of 0.91, which was slightly higher on the same corpus. Another effort on the Korean newswire documents taken similar approach to ours, and they reported their best F-score as 0.87 [13]. However these results are clearly bound to the corpus and the evaluation measures, therefore it would be difficult to compare performance scores directly from the one to another.

While both precision and recall equally contributed to the F-score, the recall was slightly higher than the precision. For the high recall, relatively small size of vocabulary in the text seemed to affect to the result. We presume that in the clinical setting, the requirement for the clarity of the text consequently restricted the vocabulary of temporal expression as well as the vocabulary size for the expression. However, for the slightly low precision, the false positives from recognizing numeric measurements as temporal expressions were still main cause of the error. Another reason for the slightly

Table 1. The experiment result of the proposed system

	GS	SF	TP	PP	PR	FS
Macro averaged	805	827	740	0.894	0.890	0.892
Micro averaged	805	827	740	0.895	0.919	0.907

GS: gold standard, SF: system found, TP: true positive, PP: phrase-level precision, PR: phrase-level recall, FS: f1-score.

Table 2. Result comparison with the validation of temporal expression and without the validation

	GS	SF	TP	PP	PR	FS
With validator	805	827	740	0.895	0.919	0.907
Without validator	805	938	740	0.788	0.919	0.849
	+/-			+13.6%	0.0%	+6.8%

GS: gold standard, SF: system found, TP: true positive, PP: phrase-level precision, PR: phrase-level recall, FS: f1-score.

http://dx.doi.org/10.4258/hir.2011.17.3.150

low precision was the duplicated count of errors in the evaluation. When the expression "9/28 '00" observed in the text, the human annotator recognized as one expression while our system recognized as two separated expression. In this case, two false-positives were counted because those two expressions were falsely recognized, and one false-negative were counted because the whole expression was not recognized by the system.

Although it is the initial approach for processing Korean clinical narrative, the result was promising. Further study will be focused to the normalization of the temporal expression, development of new approach for the more accurate recognition based on the experience and insight of the problem obtained from this study.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0028079).

References

1. Deshpande AM, Brandt C, Nadkarni PM. Temporal query of attribute-value patient data: utilizing the constraints of clinical studies. Int J Med Inform 2003; 70: 59-77.
2. Shahar Y, Combi C. Timing is everything. Time-oriented clinical information systems. West J Med 1998; 168: 105-113.
3. Bui AA, Taira RK, El-Saden S, Dordoni A, Aberle DR. Automated medical problem list generation: towards a patient timeline. Stud Health Technol Inform 2004; 107(Pt 1): 587-591.
4. Allen RB. A focus-context browser for multiple timelines. In: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries; 2005 Jun 7-11; Denver, CO. p260-261.
5. Aliferis CF, Cooper GF, Pollack ME, Buchanan BG, Wagner MM. Representing and developing temporally abstracted knowledge as a means towards facilitating time modeling in medical decision-support systems. Comput Biol Med 1997; 27: 411-434.
6. Verhagen M, Mani I, Sauri R, Littman J, Knippen R, Jang SB, Rumshisky A, Phillips J, Pustejovsky J. Automating temporal annotation with TARSQI. Stroudsburg, PA: Association for Computational Linguistics; 2005.
7. Ahn D, Adafre SF, de Rijke M. Extracting temporal information from open domain text: a comparative exploration. J Digit Inf Manag 2005; 3: 14-20.
8. Zhou L, Parsons S, Hripcsak G. The evaluation of a temporal reasoning system in processing clinical discharge summaries. J Am Med Inform Assoc 2008; 15: 99-106.
9. Kim H. Korean national corpus in the 21st century Sejong project [Internet]. [cited at 2011 Aug 16]. Available from: http://tokuteicorpus.jp/result/pdf/2006_006.pdf.
10. Stolcke A. SRILM-an extensible language modeling toolkit. In: Proceedings of 7th International Conference on Spoken Language Processing (ICSLP); 2002 Sep 16-20; Denver, CO. p901-904.
11. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010; 17: 514-518.
12. Hacioglu K, Chen Y, Douglas B. Automatic time expression labeling for English and Chinese text. In: CICLing 2005: Sixth International Conference on Intelligent Text Processing and Computational Linguistics; 2005 Feb 13-19; Mexico City. p548-559.
13. Jang SB, Baldwin J, Mani I. Automatic TIMEX2 tagging of Korean news. ACM Trans Asian Lang Inf Process 2004; 3: 51-65.