http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (18): 4441-4446, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.4441.4446
© 2013 Asian Network for Scientific Information

Extracting Medical Records with Hierarchical Information Extraction Method

[1]Wenhao Zhu, [1]Chaoyou Ju, [1]Wei Xu, [2]Jiaoxiong Xia and [1]Li Fu
[1]School of Computer Engineering and Science, Shanghai University,
[2]Information Centre, Shanghai Municipal Education Commission, Shanghai, China

Abstract: Traditional Chinese Medicine (TCM) has a very long history in China. As a part of Chinese culture heritage, clinical TCM records were preserved in TCM books. With the rapid development of digitization movement, a lot of these books are being digitized and it will be very useful if the medical records can be extracted as structural information. However, the content of TCM records is in old Chinese language and has diverse written styles as they are accomplished by different authors. Hence, it's difficult to extract these records by a general one-step approach. In this paper, we present a hierarchical information extraction method that extracts medical records in a multi-level way. Corresponding algorithms are designed according to different information level respectively so that not only the detailed textual features, such as written style and printing format but also the relations between these information are taken into account during the process of extraction. We verify our approach with TCM books which are in old Chinese language and are hard to process with normal natural language processing techniques. The experiment shows that our approach achieves a good performance for most of the test books and can be applied for other similar tasks.

Key words: Information extraction, traditional chinese medicine, medical records

INTRODUCTION

Traditional Chinese Medicine (TCM) has a continuous five-thousand years history and has its own theory and clinical experience in dealing with diseases. It has unique characteristics on comparing with western medicine. As a part of Chinese culture heritage, many TCM books are preserved and a lot of these books have special and important reference value. With the rapid development of digitization movement, it will be very useful if medical records can be extracted from these TCM books.

Information extraction is a method that can extract structured information out of semi-structured or unstructured text data. As a branch of natural language processing and artificial intelligence, the research on information extraction has been carried out since early 1980s driven by Message Understanding Conference (MUC) and Automatic Content Extraction (ACE) (Grishman and Sundheim, 1996). At present, information extraction technology has been emphasized by diverse research fields, such as database, WWW, knowledge discovery, semantic web and information retrieval. The development of information extraction evolved from manually constructing extraction rules to automatically generating rules according to samples and then to using statistical machine learning methods with rules.

Medical records extraction is a hot application issue of information extraction. Generally speaking, the task of medical record extraction is to acquire record units and each unit contains necessary information about one medical treatment case. Usually, the written style and professional language used in medical texts have their special characters and that makes medical record extraction heavily domain related.

However, the content of TCM records is in old Chinese language, which is hard to process with normal natural language processing techniques. Moreover, the TCM documents contain multiple and diverse written styles as they were accomplished by different authors. Hence, it is difficult to adopt general information extraction approaches on the problems of TCM extraction.

On the other hand, although presented in old Chinese language, medical records in TCM books can be presented with a certain information structure. The fields of this information structure usually include patient information, symptoms, diagnosis (or called dialectic analysis in TCM) and therapeutic process etc. Besides, there is also some irrelevant information that may not be very useful for a TCM information system but could be useful to improve extraction performance.

In this study, we propose a hierarchical information extraction method that extracts medical records in a multi-level way. We design different

Corresponding Author: Wenhao Zhu, School of Computer Engineering and Science, Shanghai University, Shanghai, China

algorithms at different semantic or information level to achieve a precise extraction performance. It's a relative simple method that not only focuses on detailed record information, such as symptoms and prescriptions but also the relations between the information.

To do that, we design our methods as a three-step procedure. First, our algorithm operates in paragraph level (called paragraph pre-labeling). At this level, text paragraphs are roughly marked as one of the following tags: irrelevant information, basic information, multiclass information, symptom information, prescription information and therapy information. After that, we try to identify medical record units (called record unit identification) according to the pre-labeled paragraphs with Conditional Random Fields (CRFs), which use text features and sequences of states to predict unknown sequence. Since the relations between paragraphs as paragraph sequence are taken into account, some un-labeled or mis-labeled paragraphs in previous step are labeled or corrected according to the structure sequence of record unit. The last step is called detailed information extraction. That is, based on the result of record unit recognition, specific extraction methods are employed to deal with corresponding labeled paragraphs. For example we can use a personal information extractor on paragraphs with label of 'basic information' to get the information of a patient.

The rest of this paper is organized as follows. In Section 2, we discuss the related works. Section 3 presents the hierarchical information extraction method. And we show experimental results in Section 4. At last, a conclusion is given in Section 5.

RELATED WORK

Medical information extraction: With the technical development of information extraction, a lot of work has been carried out to extract medical records from digitized content. Marciniak *et al.* (2004) introduced a medical records extraction method which extracts single pieces of information using SProUT (a general-purpose IE platform) and then, externally merge the results in order to obtain a detailed and formalized description. Rani *et al.* (2011) used two different methods to implement medical information extraction. Another possible solution was to use WordNet (Miller, 1995) to build relation semantic web between syntactic units. For these kind of methods, the key idea was to build an information extraction model for a specific domain.

Besides, there are also some works on TCM record extraction. Hu *et al.* (2008) used semantic annotations as extraction rules of medical information to solve the problem of semantic absence. Zhang (2009) employed text classification for TCM record extraction. He used Meta-Bootstrapping algorithm with a model structure to extract information terms.

Hierarchical information extraction: Currently, the application of information extraction is still quite limited. Besides from the domain adaptation problem, it's very difficult to extract from free texts due to language diversity. However, since the information existing in the text is usually correlated in most extraction tasks, it is possible to develop an algorithm that focus on not only the information itself but also the relations which are in a higher level. Hierarchical information extraction is one of the multi-level approaches.

Surdeanu *et al.* (2010) proposed a hierarchical method in legal information extraction. Their method labeled data as relevant text and entity reference of the specific domain. Based on data labeling, they showed that complex information extraction systems could be built and trained using hierarchical and partially-labeled data. The study on extracting semi-structured information from the Internet was carried out by Yu (2005). In order to deal with complex context and noisy data, he proposed the idea of cascaded double-layer for information extraction, which effectively improved the performance of extraction. Wu (2004) designed a hierarchical BBS (Bulletin Board System) browse system. This system implemented a quick location and extraction method by dividing keywords into different levels. Bounhas and Slimani (2010) proposed a top-down indexing method which valued the greatest importance to terms that appeared in the head nodes of the document. Terms were weighted according to their positions in the hierarchical structure of the document. Once the documents were indexed, logical relationships between their fragments were mined to build a contextual network of terms.

Although, there are some works on medical record extraction as well as hierarchical information extraction, most of the works discussed above are focused on either specific application domain or non-medical related problem. In this paper, when the problem comes to extract medical record from TCM corpus, it is reasonable to believe that a hierarchical approach may come up with a better performance despite the fact that natural language processing may not applicable.

HIERARCHICAL INFORMATION EXTRACTION

Figure 1 shows the general technical process of our hierarchical approach. The input text data is from TCM books (shown in the bottom). The fields need to be

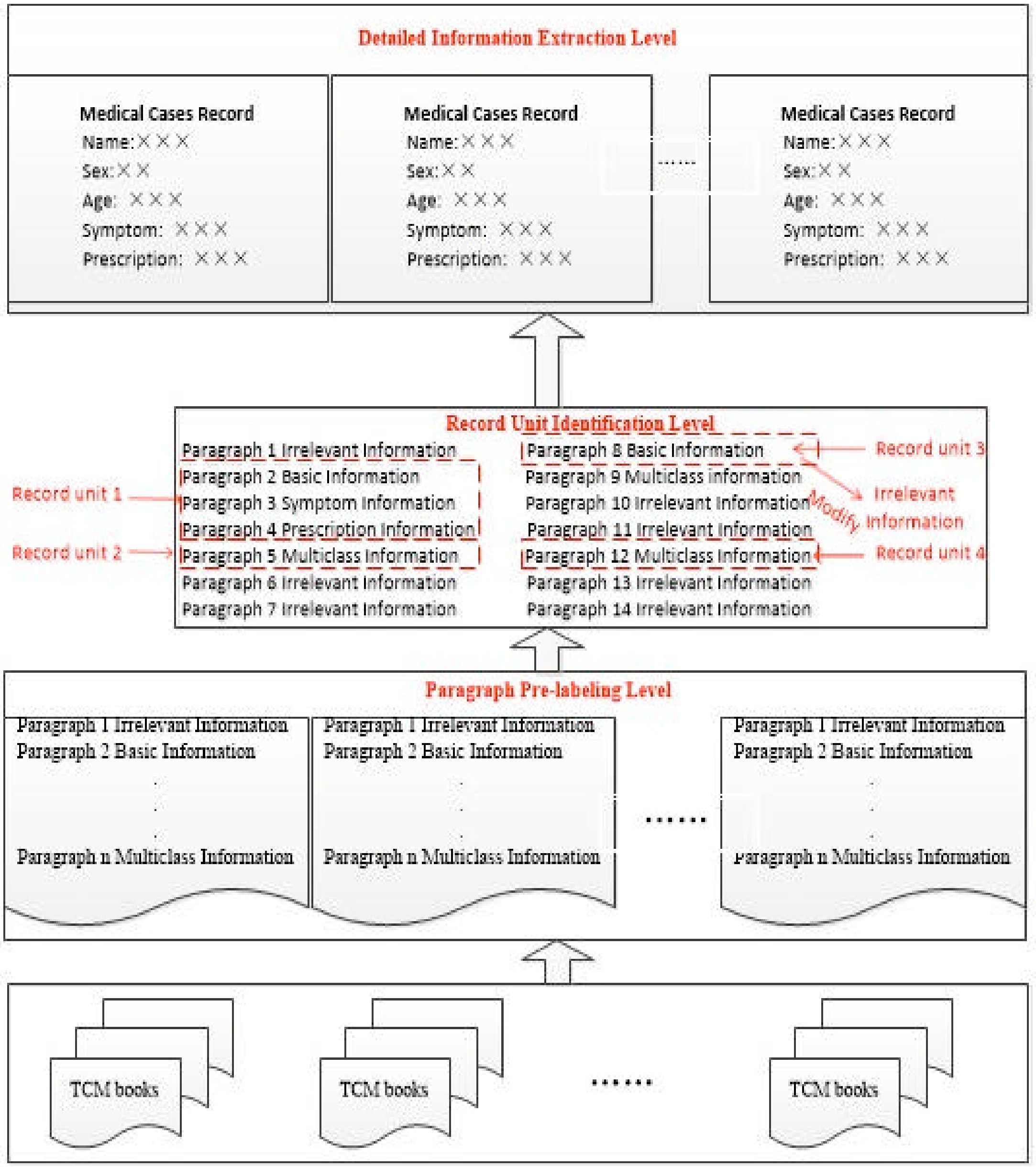
Detailed Information Extraction Level
Medical Cases Record
Name:×××
Sex:××
Age: ×××
Symptom: ×××
Prescription: ×××
......
Record Unit Identification Level
Paragraph 1 Irrelevant Information
Paragraph 2 Basic Information
Paragraph 3 Symptom Information
Paragraph 4 Prescription Information
Paragraph 5 Multiclass Information
Paragraph 6 Irrelevant Information
Paragraph 7 Irrelevant Information
Paragraph 8 Basic Information
Paragraph 9 Multiclass information
Paragraph 10 Irrelevant Information
Paragraph 11 Irrelevant Information
Paragraph 12 Multiclass Information
Paragraph 13 Irrelevant Information
Paragraph 14 Irrelevant Information
Record unit 1
Record unit 2
Record unit 3
Record unit 4
Modify
Irrelevant Information
Paragraph Pre-labeling Level
Paragraph 1 Irrelevant Information
Paragraph 2 Basic Information
Paragraph n Multiclass Information
......
TCM books
TCM books
TCM books
......

Fig. 1: General technical process

extracted include patient basic information, patient symptoms, diagnosis (or called dialectic analysis in TCM) and therapeutic process. We use paragraph pre-labeling first to provide initial labels and then mark or modify labels through record unit identification by the assumption that the record units existing in books are organized in a relative similar way. Finally, we use specific extractors to perform detailed information extraction according to the labels of every paragraph.

Paragraph pre-labeling: The first step of our approach is paragraph pre-labeling. Because the result of this step is to provide initial labels for next steps and some of unlabeled or mis-labeled paragraphs can be labeled or corrected during the record unit identification, paragraph pre-labeling needs to be as precise as possible. For paragraphs that are hard to determine, we can simply mark them as irrelevant or multi-class as they can be dealt in next steps. In this paper, we adopt gain ratio based decision tree to ensure the pre-labeling performance.

Figure 2 is the detailed processing flow of paragraph pre-labeling. First, word segmentation and tagging module tools are used to filter the stop-word (characters that marks the end of phrases, sentences and paragraphs). Subsequently, features of medical records (such as the frequency of keywords) are calculated and parameterized according to text content. After that, these parameterized features are imported with decision tree. Finally,paragraphs are labeled as one of follow tags: Irrelevant information, basic information, multiclass information, symptom information, prescription information and result information. Besides, post-pruning algorithm is combined with the original decision tree algorithm to overcome the over-fitting problem from gain ratio.

Record unit identification: Record unit identification operates at record level or semantic level if we treat the relations between record information as semantic relations. In this step, record units are identified by the assumption that the record units existing in books are presented in a relative similar way. That is, the written style, information sequence and even the length of each record will usually seems similar as long as the units of medical record are from the same book or textual source.

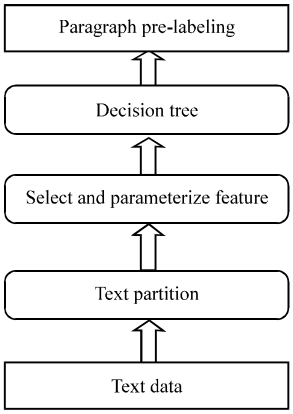

Fig. 2: Technical process of paragraph pre-labeling

Figure 3 is the processing flow of record unit identification. We identify medical record units according to the pre-labeled paragraphs with Conditional Random Fields (CRFs). CRFs is a typical discriminative model proposed by Lafferty *et al*. (2001), which models the target sequence based on the observation sequence. By regarding paragraph label order as state sequence, which usually starts with basic patient information and ends with the next paragraph of basic patient information (belongs to another record unit), we use CRFs to predict or correct paragraph labels sequentially. Besides, if a pre-labeled tag is inconsistent with the result CRFs predicted label, an additional decision procedure is employed according to the completeness of the current record unit.

After record units are identified, we can then use specific extractor (or information extraction method) to perform detailed information extraction on specific paragraphs. Figure 4 is a possible example of detailed

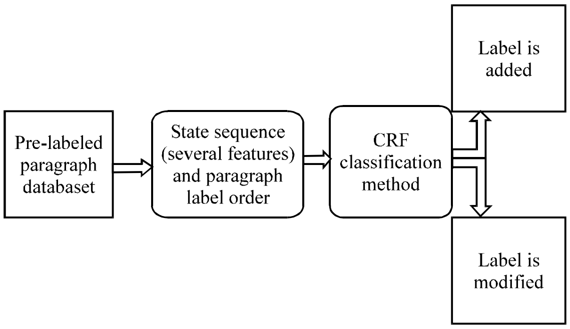
Pre-labeled paragraph databaset
State sequence (several features) and paragraph label order
CRF classification method
Label is added
Label is modified

Fig. 3: Technical process of record unit identification detailed information extraction

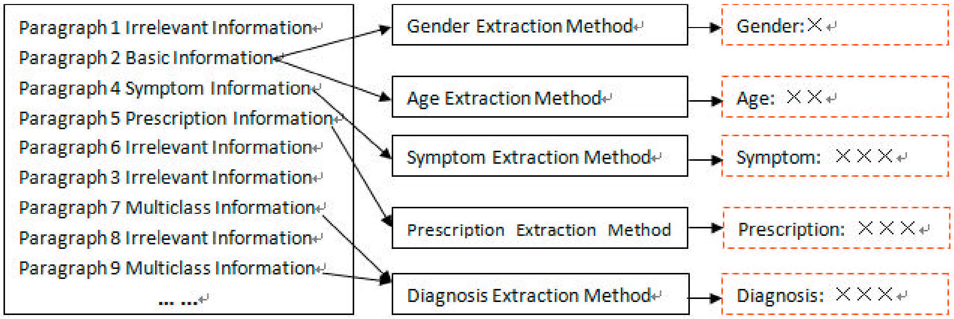
Paragraph 1 Irrelevant Information↵
Paragraph 2 Basic Information↵
Paragraph 4 Symptom Information↵
Paragraph 5 Prescription Information↵
Paragraph 6 Irrelevant Information↵
Paragraph 3 Irrelevant Information↵
Paragraph 7 Multiclass Information↵
Paragraph 8 Irrelevant Information↵
Paragraph 9 Multiclass Information↵
... ...↵
Gender Extraction Method↵
Age Extraction Method↵
Symptom Extraction Method↵
Prescription Extraction Method
Diagnosis Extraction Method↵
Gender:×↵
Age: ××↵
Symptom: ×××↵
Prescription: ×××↵
Diagnosis: ×××↵

Fig. 4: A possible example of detailed extraction

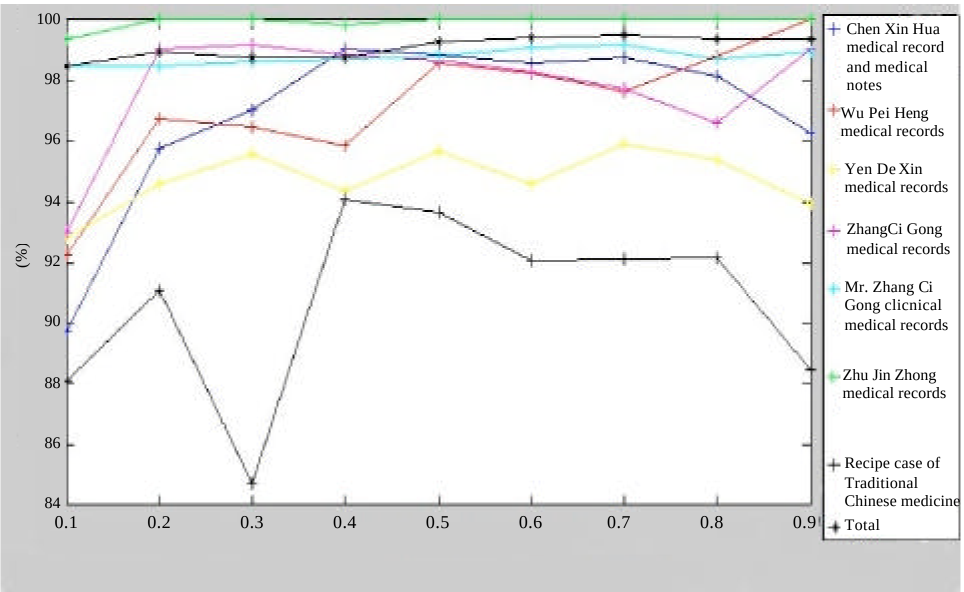

Fig. 5: Precision of paragraph classification

information extraction for one record unit. Note that the extraction method applied on the multi-class paragraphs is not decided until other paragraphs have been processed with their corresponding extractors. Because the information fields of each unit are relatively fixed in the same book, we can use the extractors that have not yet been used to extract the missing information for the record unit. For example, suppose gender, age, symptom and prescription have already been extracted according to the paragraph label, while diagnosis is missing, we can use diagnosis extraction method to get diagnosis information with un-processed paragraphs which are probably labeled as multi-class in our experiment.

EXPERIMENT

The experimental data come from 7 different TCM books. To show the detailed performance of each step (corresponding to each extract level), we divide the experiment as three parts: paragraph pre-labeling, record unit identification and integrated extraction result with detailed information extraction. Note that the extraction result of each step is not compared with the integrated result. This is because these steps are merely working flows rather than intermediate procedure.

Experiment result: The precision of paragraph pre-labeling is shown in Fig. 5. This part operates in paragraph level. X axis expresses the ratio of training data and test data in this figure. The values of Y axis represent the precision of classification. As can be seen from the figure, the precision is always greater than 0.84.

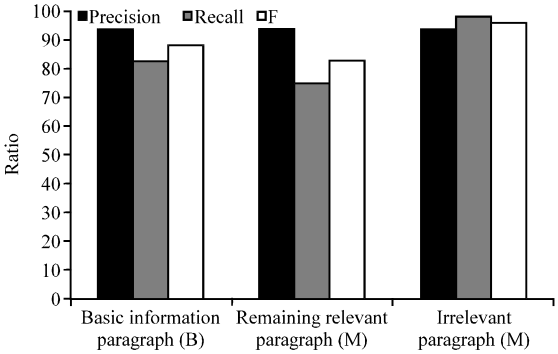

Fig. 6: Effect of record unit identification

The effect of record unit identification is shown in Fig. 6 (Record unit starts with the basic information). This part operates at record level. In this figure, X axis expresses basic information paragraph of the record unit, remaining relevant paragraphs and irrelevant paragraphs respectively. The values of Y axis represent precision, recall rate and F-score. As we can see from the figure above, the method of record unit identification can receive better classification effect.

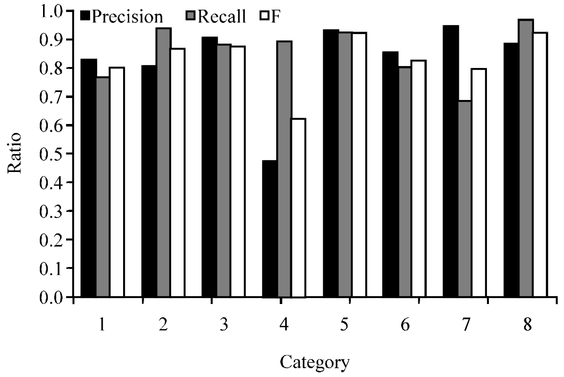

Fig. 7: Effect Of detailed information extraction

The effect of detailed information extraction is shown in Fig. 7. In this figure, X axis is the category of TCM books. The value of Y axis represents precision, recall rate and F-score. As we can see from the figure above, the precision of the fourth book (refer to "Mr. Zhang Ci Gong clinical medical records") is less than 0.5 for two reasons. One is the poor optical character recognition (OCR). The other is that the result represents all paragraphs integrated extraction performance of one book rather than one paragraph or record unit. Others are more than 0.8. The recall rate is near 0.8.

CONCLUSION

In this study, we present a hierarchical information extraction method that extracts medical records in a multi-level way. The effectiveness and correctness of medical records extraction with hierarchical information extraction method are validated through the experiment. Based on the analysis and sum-up of the research in this topic, the method can go further in the following aspects. First, this research is based on the domain of TCM, we can analyze and summerize general features to enhance versatility of the method. Second, the extracted results may be greatly affected because of poor optical character recognition.

ACKNOWLEDGMENT

The study of this study is supported by National Natural Science Foundation of China (No. 61303097), Shanghai Leading Academic Discipline Project (J50103), Ph.D. Programs Foundation of Ministry of Education of China (20123108120026) and Shanghai Municipal Science and Technology Commission (11510500300).

REFERENCES

Miller, G.A., 1995. WordNet: A lexical database for English. Commun. ACM, 38: 39-41.

Hu, X.Q., C.L. Zhou and S.Z. Li, 2008. Construction of ancient cases database and research on data processing. Intell. J., 27: 127-129.

Bounhas, I. and Y. Slimani, 2010. A hierarchical approach for semi-structured document indexing and terminology extraction. Proceedings of the International Conference on Information Retrieval and Knowledge Management, March 17-18, 2010, Shah Alam, Selangor, pp: 315-320.

Marciniak, M., A. Mykowiecka, A. Kupsc and J. Piskorski, 2004. Intelligent content extraction from polish medical reports. Proceedings of the 2nd International Workshop on Intelligent Media Technology for Communicative Intelligence, September 13-14, 2004, Warsaw, Poland, pp: 68-78.

Surdeanu, M., R. Nallapati and C. Manning, 2010. Legal claim identification: Information extraction with hierarchically labeled data. Proceedings of the LREC Workshop on the Semantic Processing of Legal Texts, May 17-23, 2010, Malta, pp: 22-29.

Rani, P., R. Reddy, D. Mathur, S. Bandyopadhyay and A. Laha, 2011. Compositional information extraction methodology from medical reports. Proceedings of the 16th International Conference on Database Systems for Advanced Applications, April 22-25, 2011, Hong Kong, China, pp: 400-412.

Grishman, R. and B. Sundheim, 1996. Message understanding conference-6: A brief history. Proceedings of the 16th Conference on Computational Linguistics, August 5-9, 1996, Copenhagen, Denmark, pp: 466-471.

Wu, X.Y., 2004. Applications of hierarchical keyword extraction and automated text classification in bulletin board system. Master's Thesis, Shanghai Jiaotong University, China.

Yu, K., 2005. Research on semi-structured information extraction of the internet. Ph.D. Thesis, University of Science and Technology of China, Anhui, China.

Zhang, Y.B., 2009. Research on data processing for traditional Chinese medicine cases. Master's Thesis, Nanjing University of Science and Technology, Nanjing, China.

Lafferty, J.D., A. McCallum and F.C.N. Pereira, 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning,. June 28-July 1, 2001,. Williamstown, MA, USA., pp: 282-289.