Compositional Information Extraction Methodology from Medical Reports

Pratibha Rani[1], Raghunath Reddy[1], Devika Mathur[2], Subhadip Bandyopadhyay[2], and Arijit Laha[2]

1. International Institute of Information Technology, Hyderabad
2. Infosys Technologies Ltd., SETLabs, Hyderabad
{pratibha_rani,raghunath_r}@research.iiit.ac.in
{subhadip_b,devika_mathur,arijit_laha}@infosys.com

Abstract. *Currently health care industry is undergoing a huge expansion in different aspects. Advances in Clinical Informatics (CI) are an important part of this expansion process. One of the goals of CI is to apply Information Technology for better patient care service provision through two major applications namely electronic health care data management and information extraction from medical documents. In this paper we focus on the second application. For better management and fruitful use of information, it is necessary to contextually segregate important/relevant information buried in a huge corpus of unstructured texts. Hence Information Extraction (IE) from unstructured texts becomes a key technology in CI that deals with different sub-topics like extraction of biomedical entity and relations, passage/paragraph level information extraction, ontological study of diseases and treatments, summarization and topic identification etc. Though literature is promising for different IE tasks for individual topics, availability of an integrated approach for contextually relevant IE from medical documents is not apparent enough. To this end, we propose a compositional approach using integration of contextually (domain specific) constructed IE modules to improve knowledge support for patient care activity. The input to this composite system is free format medical case reports containing stage wise information corresponding to the evolution path of a patient care activity. The output is a compilation of various types of extracted information organized under different tags like past medical history, sign/symptoms, test and test results, diseases, treatment and follow up. The outcome is aimed to help the health care professionals in exploring a large corpus of medical case-studies and selecting only relevant component level information according to need/interest.*

Keywords: Information Extraction, Medical document mining, Health care application, Clinical Informatics

1 Introduction

Clinical Informatics (CI) is a recent field of IT application research emphasizing better quality of patient care in simultaneity with cost optimization. This in

turn promises a huge scope of business application in health care industry. The core technology behind this lies in the domain of electronic health care data management and information extraction from unstructured documents, the two parallel mainstreams in CI. The resulting applications induce better decision making in different contexts like better treatment provision, enhancing quality of life of patients and so on.

In this paper we consider the IE field and present a compositional approach for information extraction from free format text related to patient care process. Our aim is to extract information relevant to different context in the form of passage/collection of sentences from documents and present them in a composed, self contained format. This approach has one essential generic concept; a document creation is an outcome of evolution of a compound activity in a specific domain. From initiation to completion the compound activity is viewed in terms of granules of interlinked sub-activities at different intermediate stages creating interdependent contextual information packets as output. These contextual information are distributed along the corresponding document(s) in an entangled manner. An illustration of this concept in a patient care process is given in Fig. 1 as an ordered activity network. The document creation is actually carried out following the underlying activity network in patient care domain. By relevant extraction of information we emphasize the fact that our approach will extract and organize information pertinent to these individual contexts.

Information extraction from medical documents has been addressed mostly from discrete perspectives where the interest is usually on a few specific components. The major challenges in building a holistic approach for relevant information extraction from texts generated in patient care process are non explicitness, repetition of information across the document in varied expression and overlapping of information belonging to different implicitly expressed contexts. The bottom up view of a document creation through compilation of information artifacts generated from integrated sub-activities, as emphasized in this paper, helps to overcome this problem. Along the same line of thought, we propose individual modules for information extraction from each class (sign/symptom, past medical history etc.) and thus the whole process can be viewed as an integrated system. The extracted fragments of information from different classes are ultimately compiled to make a complete structure.

The rest of the paper is organized as follows. Section 2 discusses the motivation and section 3 presents related studies in context to this and similar problems. The methodologies and proposed algorithms are discussed in section 4. In section 5 we present experimental study and results along with discussion on the computational aspect of this problem. Finally section 6 concludes the article with some comments on our focus and nature of the solution.

2 Motivation

Let us consider the representation of a patient care activity in terms of activity flow (Fig. 1) which is the motivation pivoting subsequent development of

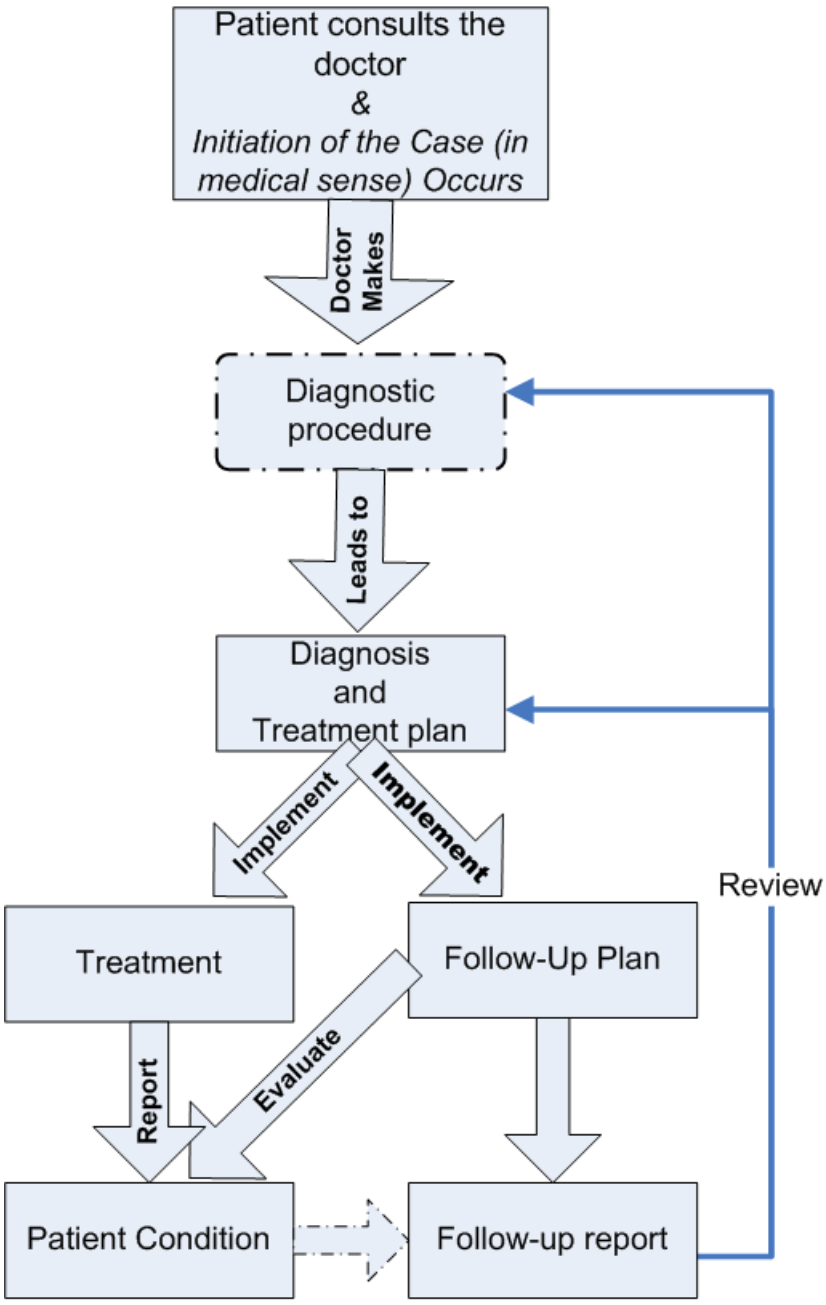

Fig. 1. Patient care process flow

concepts and discussions in this paper. As depicted in the figure, we perceive a patient care process through different sub activities namely collection of past medical history, observing sign/symptom, suggesting tests, observing test results, confirming a diagnosis, prescribing treatment and pursuing a follow up. Thus different types of information like symptoms, medical history etc. are generated contextually in the execution of the sub-activities. The relevant information in context to a physician's interest is actually contained in six different medical entities namely past medical history, sign/symptom, test and results, diagnosis, treatment and follow up. These entities can be assumed to represent six different classes of information with some class specific or contextual characteristics which we explore and exploit to construct heuristic extraction rules. For example, the vocabulary, semantics and sentence format corresponding to sub-activities symptoms collection and medical test result composition are markedly different. The first type is a mixture of deep semantics related to feelings and observations on clinical events often forming different types of regular expressions, where as the second type is more prominent with typical medical vocabulary.

With the process flow, information piles up across the (sub)activity layers to form a medical case report. This fragmented view of a document through different contexts is the key motivation for compositional approach. Hence for extracting relevant information from a medical document we can map the relevance of the information artifacts with the underlying sub-activities and hence can form clear guidelines on the target set. Thus an approach that is composed of differentiated extraction task for individual types of information seems to be a natural choice.

From the activity point of view, a patient care process starts with patient coming to a doctor and continues through subsequent stages viz., diagnosis, treatment, follow up, review. Intermediately it iterates in the diagnosis-treatment-follow up-review-diagnosis or treatment-follow up-review-treatment cycle (or a cycle combining part/whole from these two cycles) until a conclusion (cure, patient quitting treatment or death of the patient) is reached. Thus the evolution of a patient care process described in terms of combination of sub activities, in perception of a physician's context, is natural and seems to be justified enough to work upon. The relevant information classes across different research reports created by physicians thus remain the same. There may be ornamental changes in presentation of the corresponding texts but the class specific characters stay close to the information class which can be exploited for information extraction.

It can be noted that different contexts may correspond to different types of information as relevant and hence the construction of the extraction rules will change. For example, from a pharmacists point of view, the medicines prescribed and the corresponding chemical groups might be of more interest. Here the consideration of the contextual nature of information within a document and with respect to a user is the differentiator of our extraction approach. Depending on the context a user (e.g. a physician) might be interested in specific information, like finding suitable tests given a set of symptoms or the set of treatments given

a disease. The proposed approach of information extraction can be applied easily and efficiently to address such needs.

Thus any document arising from a patient care activity has the inherent structure consisting of six relevant information class with class specific characteristics which we intend to exploit for information extraction. This is explained in the subsequent discussions. It can be noted that the diagnostic procedure itself is made up of a complex flow of activities that we have not considered separately.

3 Related Work

Information extraction from medical document is a long standing area of research addressed by a mixture of research communities during past few decades. An excellent survey is available in [6]. Among the relevant papers, [17] uses a SVM based supervised model to annotate unseen terms in new texts and contexts based on manually annotated terms by domain experts. [21] presents an approximate dictionary-based biological concept extraction method where the basic idea is to capture the significant words rather than all words of a concept which is more related to biomedical field rather than a patient care scenario. In [20] document retrieval is done on the basis of concepts and their relations. The basic difference between our approach and these studies is that our focus is on relevant part within a document rather than the document as a whole.

The study in [8] has some similarity with our thought process but they explore more in terms of hidden relation extraction using conditional random field based approach. [9] uses a graphical model based on extension of the basic Latent Dirichlet Allocation framework for indexing PubMed abstracts with terminological concepts from ontology. Similar type of study consisting of sub-topic extraction is considered in [11]. Study conducted in [13] can be identified as a part of the whole scenario that we have considered here. In [13] a NLP application is designed to extract medical problems from narrative texts in clinical documents that come from a patient's electronic medical record.

[15] proposes a biological ontology (provided in UML S) based technique for extracting summarization of texts obtained from BioMed Central. [19] implements a medical Information Extraction (MedIE) system that extracts a variety of information from clinical medical records. Taking the help of section headings they perform ontology based extraction of medical terms, graph-based extraction of relations using link-grammar parser and text classification using ID3-based decision tree. [10] presents a framework for patient data extraction along the line of methodology describes in [19] with an automated storing process in a relational database.

It can be noticed that at micro-level the output of our approach is a collection of sentences belonging to different types of information related to the sub-activities in the process flow of patient care. Application for executing similar (micro-level) task, perceived mainly as passage level information extraction, is well discussed in literature, viz., [11, 12, 14, 18]. But the integrated approach to

combine them for a single purpose of IE from patient care data is little discussed in the existing literature which is the focus of this study and the value addition of this article. Also note that our integrated approach can harness any such technological/methodological advancement in re-usability context and increase the net value addition.

4 Compositional Information Extraction Method

The underlying activity flow of a patient care process illustrated in Fig. 1 is the motivation behind the compositional information extraction approach. As we have noticed, the activities involved in different levels of a patient care process generate different classes of information which are of interest to a typical user (e.g., a physician). Each of the class has some unique characteristic structure, type of key word and phrase, semantics, vocabulary and so on. Due to these profound interclass differences, a single holistic approach is not an appropriate way to address the problem. Instead we plan for differentiated modules, one for each class of information and combine them on a common platform for the ultimate execution. The information extraction process adopted in this paper is an integrated system of three parallel and mutually interacting building blocks:

1. Regular expression based pattern matching.
2. Dictionary based lookup and matching.
3. Heuristic based passage extraction algorithms.

We extract patient related information like age, gender etc. using regular expressions. Dictionary based lookup along with regular expression based pattern matching is used for identifying medical tests and test results. Regular expressions are required to identify test result related texts which contain numerical information separated by measuring units like 130/80 $mmHg$, 9.84 gm/dl etc.

Dictionary based lookup is the common approach used for identifying medical entities like disease, diagnosis, drugs, treatment, test and results, sign/symptoms and follow-up. A category specific dictionary contains a list of words related to the corresponding medical entity which can always be improved through domain expert intervention and hence the extraction efficiency can be improved as well. It is important to note that test, sign/symptom and past medical history information are many a times overlapping which need some methodology to differentiate. So by carefully analyzing the available case reports we develop heuristics to handle this.

In the next subsections we discuss information extraction methods for the class past medical history and follow up in detail as there are some typical complexities to handle.

4.1 Extracting Past Medical History Passages

Past medical history related information in the type of documents we have considered has some inherent characteristic which restricts the direct applicability

of some reported approaches like in [7] and [16]. The authors of [7] propose a robust corpus-based approach for temporal analysis of medical discharge summaries.This learning based approach requires large trained corpus and will be computationally very expensive when we just need to extract past medical history. The authors of [16] investigate four types of information found in clinical text that helps in understanding what textual features can be used in extraction of past medical history which are typically not much prominent in the research report format.

One of the basic problem in the documents we consider is that the semantics in these documents do not explicitly express past event related structure since the documents are written as reports of events observed some time back and hence the current context is also expressed in past format. Also precise chronological statement is not available; even if it is there it is mostly not in an explicit hierarchical order. Thus the usual approaches that rely on graphical representation of temporal events etc. have a limited or no scope of applicability. We use heuristic based approach to tackle this problem.

After analyzing the case reports we found that past medical history usually appears in first half of the case report. Also it may be as a group of sentences in the beginning or embedded within the case report. We use the spatial order of sentences present in the report to extract the past medical history. The narration of the case is assumed to be the present (on current state of patient) and anything cited before in time will be past history.

We use simple *Allen's temporal logic* [7] to find whether one sentence comes before, after or has no relation with other sentence to identify past medical history sentences and present sentences. We then identify frequent keywords found in both type of sentences and use them in extracting passages related to past medical history.

---

history pattern = {for several years, was on therapy, ago, past, history, year previously, no previous, years previously, many years, when aged, week before, months earlier, years earlier, previous year, before admission, prior to admission, prior to presentation}

present pattern = {while, treated, initial, demonstrated, showed, confirmed, investigations, reveal, complained, initially, given, treatment, exam, diagnostics, presentation, received, arrived, admission, now, normal, diagnosed, was admitted, presented to, on along, presented with, admitted with, admittance, upon arrival, indicative, indicated, discharge, on arrival, yet, so far, shortly, presently, recently, follow up, meanwhile, within, physical examination, on physical examination}

Fig. 2. Patterns of History and Present text phrases.

FindWordSeq Algorithm:

1. Find frequency of each word in the text /*store it in a Hashmap*/
2. Repeat step 3 for window size k= 1, 2,3
3. While (not end of text)
   (a) Use a sliding window of size k
   (b) If (all the words in the window are frequent) /*use Hashmap values*/
      – Output the word sequence
   (c) Read new text in window

Fig. 3. Algorithm to find Frequent Word Sequences.

Identifying Frequent Word Patterns:

1. Use *Allen's Temporal Logic* to make two set of files – one set containing sentences belonging to past medical history and other set containing non past medical history sentences (present).
2. Extract frequent single words from the two set of files using Mafia tool.
3. Use algorithm *FindWordSeq* to find frequent continuous word sequences from the two set of files.
4. Remove duplicates from both the sets.
5. Remove common phrases present in both the sets.
6. Label the frequent text phrases of past medical history set files as *history pattern* and the frequent text phrases of non past medical history set files as *present pattern*.

Fig. 4. Identifying frequent word patterns of Past Medical History and Present Sentences.

We use **Mafia** tool [1] along with a sliding window based algorithm *FindWordSeq* (Fig. 3) to find frequent text patterns or words found in past medical history and present sentences. Mafia tool is designed to find single frequent words which we utilize at the outset. Then after removing the duplicates we use these single frequent words in a sliding window based algorithm *FindWordSeq* to find frequent continuous sequence of words (up to size 3). Time complexity of *FindWordSeq* algorithm is $O(n)$, where n is number of words in a file. Note that choice of Mafia tool is just for the sake of scalability and open access. Otherwise frequent single word finding can be addressed independently without any difficulty. After analyzing the available documents we found that existing frequent continuous word patterns are of maximum length 3, so we set the upper limit of sliding window size to 3. We also observed that inclusion of text patterns found by *FindWordSeq* improves the performance of the past medical history extraction module.

The process of finding frequent text patterns is explained in Fig. 4. The obtained frequent text phrases in past medical history sentences (*history pattern*) and the frequent text phrases in non past medical history sentences (*present pattern*) are shown in Fig. 2. We then use a simple heuristic based algorithm *ExtractHistory* (Fig. 5) to extract past medical history passages. Using the *history pattern* and *present pattern* set this algorithm exploits the chronological sentence order present in the case report to extract the past medical history of a patient. One single scan of the text provides the required information. Time complexity of this algorithm is linear and depends on the number of sentences.

ExtractHistory Algorithm:

---

1. Mark each sentence as history category sentence or present category sentence on the basis of phrases or words in *history pattern* and *present pattern* respectively.
   (a) If a sentence contains words in *history pattern* mark it as history category sentence.
   (b) If a sentence contains *present pattern* mark it as present category sentence.
   (c) If a sentence contains words both in *history pattern* and *present pattern* mark it as history category sentence.
2. The sentence belonging to history category sentence marks the beginning of past medical history (PMH). Go on including sentences in PMH until first present category sentence is encountered. (The sentences marked in step 1 helps in identifying the beginning and end of past medical history).
3. Repeat step 2 to find past medical history sentences that are found at different places in the report.

Fig. 5. Algorithm for Extraction of Past Medical History Sentences.

4.2 Follow Up Text Passage Extraction

Two major problems of extracting follow up related information are notably less volume (in terms of sentence length and/or number of sentences) of it's description and non-explicitness of the structure. However the spatial information of it's location can help in the extraction. The follow up text passage is usually present at the end of a medical case report. We first find the most frequent keywords found in follow up. If a sentence contains these keywords then it is added to follow up passage. We also use the heuristic of including the sentences of last paragraph before *Conclusion* or *Discussion* section as follow up since by manual analysis we found them usually to be follow up sentences.

5 Experiments and Results

We use three softwares namely Eclipse [2], UIMA (Unstructured Information Management Architecture) [3] and MySQL [4] for developing the tool. Most of the coding is done in Java apart from some preprocessing tasks done using Perl. We use UIMA framework for implementing the algorithms as it provides a good platform for integrating different sub-modules for executing a complex task through combination of simpler sub-tasks. Also it provides support for important functions like entity annotation, regular expression annotation, POS tagging etc. For experimentation we use the heart disease related research reports collected from *Journal of Medical Case Reports* (a open source archive [5]). The collected data is first preprocessed to remove figures, links, references etc and then converted into simple text files. These files are supplied as input to the developed tool.

A typical medical case report is in general, but not limited to, a free text format description of a patient care event starting from past medical history related description followed by observations and discussions on symptoms/signs, medical tests and results leading to diagnosis, treatment details covering medication, surgery or different therapy and concluded by follow up. The typical features of such type of texts described in a research paper format in the *Journal of Medical Case Reports* are realized in the following aspects:

1. There are two broad sections containing the main case presentation and discussion but typically the information from different fields are entangled. One notable thing is that the follow up part is often found in the discussion part entangled with other informations not of interest to us.
2. The sequence of events describing a patient care process is not followed quite often and later issues are described first.
3. The domain dependence of the vocabulary.
4. Occurrence of numerical values with typical units and characters.
5. The whole report is written after the completion of treatment. Hence recognition of the past medical history related sentence/passage is very confusing.

We use training corpus to learn the regular expressions and other notable features to derive heuristic rules. The results of the manually tagged test corpus is compared with the system extracted results to judge the performance. We consider sentence as the unit level of comparison and compute Precision and Recall by comparing the system extracted sentences and the corresponding manual extraction. Since the information belonging to the different information classes are not distributed uniformly within and as well as between documents, a better way of performance measurements is to consider individual classes over the corpus and evaluate the performance separately for them. We use the macroscopic method of combining the results in which equal weight is given to all the samples and so, Precision and Recall values are averaged over all the test samples.

Table 1 presents the overall Precision, Recall and F measure values obtained over test corpus for classes Past Medical History, Sign/Symptom, Test and Test Results, Disease/Diagnosis, Treatment and Follow up. Note that F_1 measure

is the harmonic mean of Precision and Recall measures while F_2 gives twice weightage to Recall and $F_{0.5}$ gives twice weightage to Precision. Definition of F_2 and $F_{0.5}$ measures are given below:

$$F_2 = (1 + 2^2) \cdot Precision \cdot Recall/(Precision + 2^2 \cdot Recall)$$

$$F_{0.5} = (1 + 2^2) \cdot Precision \cdot Recall/(2^2 \cdot Precision + Recall)$$

Table 1. Performance measure values for the six information classes.

Class	Precision	Recall	F_1	F_2	$F_{0.5}$
Diagnosis	1	0.849206	0.918455	0.875614	0.965704
Signs/Symptoms	0.457173	0.412698	0.433799	0.420887	0.447528
Past Medical History	0.774376	0.588889	0.669014	0.61852	0.728485
Test and Results	0.714361	0.379107	0.49534	0.418376	0.607003
Treatment	0.5	0.39418	0.440828	0.411602	0.474522
Follow up	0.444709	0.620075	0.517951	0.574746	0.471371

5.1 Discussion

Results show excellent performance for Diagnosis and good performance for Past Medical History. For Follow up performance is average. But for Signs/Symptoms, Test and Treatment performance is less than average. One major reason of poor performance for Sign/Symptom, Test and Treatment is that corresponding dictionaries do not cover all possible cases. We expect that consulting a domain expert will significantly improve the performance for these classes. It seems that we also need to apply natural language processing techniques for Sign/Symptom class where the patient's feeling related expression characterizes the texts heavily. Same applies for Follow up also because it is expressed through deep semantics and occupies comparatively very less portion in the case reports.

An obvious point arises here regarding the comparison of our approach with others. We would like to emphasize that there is hardly any room for such a study since our approach is a holistic one and the related literature, as discussed in section 3, are mostly focused towards extraction of a few specific information type from specific kind of medical documents. While the existing literature focuses on the efficient extraction of some atypical information from a document, we concentrate on the relevant extraction of information by first defining "relevance" contextually through characterization of document creation process and user's perspective. Thus when the context shifts from a physician's interest to that of a pharmacist, the relevance might lie in a few of the medical entities, mentioned before in section 2. For example a pharmacist's interest may be in the proposed medicines and the corresponding chemical groups if such information is contained in the text.

Another point to be noted is the performance aspect of our approach. There are scopes to improve the extraction performance by adapting techniques from Natural Language Processing or exploring inter class associations of the medical entities mentioned earlier. But we want to emphasize that the ingenuity of our approach lies in the perception of "relevance" of information through document creation process where information artifacts are generated stage wise in the evolution of a compound activity. The compound activity is carried out through execution of interrelated sub activities. Our aim is to propose the construction of a system that addresses relevant information extraction in a holistic manner. Once the system construction is addressed, performance enhancement can be achieved by adapting efficient extraction processes for individual module construction. Thus the paper is not intended to explore the achievement in performance/efficiency aspect but to introduce a novel idea.

6 Conclusion

In this paper we proposed a compositional approach for extracting necessary information for six different classes namely past medical history, sign/symptom, test and results, diagnosis, treatment and follow up from a free format medical case report emerging from a patient care process. We perceive a patient care process as a sequence of multiple sub-processes and the aforesaid information classes are considered to be outcome of different sub-process(es). We proposed a module based approach where each type of information is extracted using specific module(s) and the modules are integrated to work as a single composite system. The methodology is mainly based on heuristic approach that uses regular expressions, dictionary and different types of rules that are learned from the training corpus and the associated openly available resources. It can be noted that the ingenuity of the paper is the inception of a holistic and relevant IE approach from medical case research reports which is a kind of novel in it's class and hence not much scope for comparison with existing approach and/or study on the performance measure/enhancement is conceived.

Similar idea can be ported to any domain for extraction of relevant information from a free format text. Once we perceive the process flow of the events underlying the text generation, we can define the relevance as the informational artifacts generated at the granular process levels and set the information class and execute the subsequent stages to construct a composite system.

References

1. http://himalaya-tools.sourceforge.net/Mafia/.
2. http://www.eclipse.org/.
3. http://incubator.apache.org/uima/.
4. http://www.mysql.com/.
5. http://jmedicalcasereports.com/.
6. Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. Summarization from medical documents: a survey. *Artif. Intell. Med.*, 33(2):157–177, 2005.

7. Bramsen, Philip, Pawan Deshpande, Yoong Keok Lee-and, and Regina Barzilay. Finding Temporal Order in Discharge Summaries. In *EMNLP*, 2006.
8. Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9(1):207, 2008.
9. Markus Bundschus, Mathaeus Dejori, Shipeng Yu, Volker Tresp, and Hans-Peter Kriegel. Statistical modeling of medical indexing processes for biomedical knowledge information discovery from text. In *BIOKDD08*, 2008.
10. Hyoil Han, Yoori Choi, Yoo Myung Choi, Xiaohua Zhou, and Ari D. Brooks. A Generic Framework: From Clinical Notes to Electronic Medical Records. In *CBMS '06*, pages 111–118, 2006.
11. Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 9–16, 1994.
12. Christoph Mangold. A survey and classification of semantic search approaches. *Int. J. Metadata Semant. Ontologies*, 2(1):23–34, 2007.
13. Stéphane Meystre and Peter J. Haug. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. *J. of Biomedical Informatics*, 39(6):589–599, 2006.
14. Raymond J. Mooney and Razvan C. Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explorations*, 7(1):3–10, 2005.
15. Laura Plaza Morales, Alberto Díaz Esteban, and Pablo Gervás. Concept-graph based biomedical automatic summarization using ontologies. In *TextGraphs '08*, pages 53–56, 2008.
16. Danielle L. Mowery, Henk Harkema, John N. Dowling, Jonathan L. Lustgarten, and Wendy W. Chapman. Distinguishing historical from current problems in clinical reports: which textual features help? In *BioNLP '09*, pages 10–18, 2009.
17. Koichi Takeuchi and Nigel Collier. Bio-medical entity extraction using support vector machines. *Artif. Intell. Med.*, 33(2):125–137, 2005.
18. Victoria Uren, Philipp Cimiano, José Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics*, 4(1):14–28, 2006.
19. Xiaohua Zhou, Hyoil Han, Isaac Chankai, Ann Prestrud, and Ari Brooks. Approaches to text mining for clinical medical records. In *SAC '06*, pages 235–239, 2006.
20. Xiaohua Zhou, Xiaohua Hu, Xia Lin, Hyoil Han, and Xiaodan Zhang. Relation-Based Document Retrieval for Biomedical Literature Databases. In *DASFAA*, pages 689–701, 2006.
21. Xiaohua Zhou, Xiaodan Zhang, and Xiaohua Hu. MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup. In *PRICAI*, pages 1145–1149, 2006.