CRET: A Tool for Automatic Extraction of Causal Relations[*]

Yousef Abuzir[**]

[*] Received: 16/3/2013, Accepted: 11/6/2013.

[**] Dean of Faculty of Technology and Applied Sciences, Al-Quds Open University, Palestine.

ملخص:

هناك اهتمام كبير باستخراج المعرفة، واسترجاع المعلومات بصورة تلقائية من مجموعة المصادر الإلكترونية والأدبيات الأكاديمية الموجودة على الشبكة العنكبوتية. وقد طُورت في هذا البحث أداة برمجية (سيرت CERT) لاستخراج العلاقة السببية، أي استخلاص العلاقات السببية من النصوص. والأداة البرمجية هي محلل للعلاقات تستخدم لاستخراج أنماط العلاقة من الوثائق الطبية. يُكشف عن الأنماط السببية من خلال عملية المطابقة الغامضة بين قاموس العلاقات السببية وبين عملية الكشف الجزئية عن سلسلة الأنماط في الوثائق الطبية الإلكترونية.

تخزَّن المعرفة المستخرجة كفهرسه للوثائق، وكذلك تُستخدم خلال نظام شبكي لتقديم الملاحظات والاطلاع عليها. يمكن للباحثين من إجراء الاستشارات، والتحقق من صحة قائمة المصلحات والأنماط الحديثة التي أُنشئت بصور أتوماتية من النظام. إن المساهمة الرئيسة من هذا العمل هو بناء أداة برمجية لاستخراج السبب، وتأثير الحالة، واستخراج الظروف باستخدام العلاقة الغامضة.

تعتمد طريقة استخراج السببية على اسم المستخرج والعبارات الوصفية المرتبطة بالفعل السببي (الأنماط). وقد استخدمت المعايير الكمية مثل: الدقة، والاسترجاع، ونقاط (اف) في علمية التصنيف، واستخراج الأنماط السببية لتقويم أداء النظام ونجاعته، وحسبت لتقويم النتائج لدينا. فقد أظهرت النتائج أن النظام سيرت يولد ٧٧٪ من الكلمات الرئيسة والعلاقات السببية التي عُيّنت من قبل الخبراء يدوياً.

Abstract:

There is an interest in extracting knowledge and retrieving information automatically from the current availability of a large collection of electronic resources and from the academic literature available on the Web. In this work a tool called Causal Relation Extraction Tool (CRET) has been developed to extract causal relations from texts. The tool is a Relation Parser to extract relation patterns from medical documents. The causal patterns are detected through a fuzzy matching process between the causal patterns database and partial detected string patterns in the electronic medical documents. The extracted knowledge is stored as an index for the documents and the researchers can consult the indexed databases. The main contribution of this work is a method for cause, effect and condition extraction using a fuzzy relation. The causal extraction method is based on extracted noun and adjectival phrases associated with causal verb (patterns). Quantitative matrices measurements like, Precision, recall, and F-score for the classifiers and the causal pattern extraction were used and computed to evaluate our result. The results indicate that CRET generates 77% of the keywords and the casual relations which have manually been associated by human expert.

Keywords

Causal Relations, indexing, information retrieval, and Automatic Extraction of Causal Relations.

1. INTRODUCTION:

Causality is a type of relationship between two atomic entities: cause and effect. Causality may be defined as a relationship between one phenomenon or event (A) and another (B) in which A precedes and causes B. The direction of influence and the nature of the effect are predictable and reproducible and may be empirically observed. Causal extraction from text data can be subdivided into two main approaches: extraction based on grammar patterns and extraction based on co-occurrence statistics. Extracting causal relations from unstructured text used simple grammar patterns for extracting causes and effects. These patterns consisted of causes, a causal verb, and an effect, in order to extract explicitly-stated causal relations.

In this work, we developed a tool, which is designed especially for automatic extraction of relevant knowledge from medical electronic documents. The most important part of this work is the idea of using Relation Parser to extract relation patterns from medical documents. We developed a set of patterns that specifies the different types of causal relation that can be explicitly expressed in a sentence. We call them causality patterns. CERT detected the causal patterns through a fuzzy matching process between causal patterns and partial detected string patterns in the electronic medical documents.

In section 2 we discuss related work. Section 3 gives an overview of the structure of the tool and the main components of the system. In section 4 we describe documents collection. Section 5 discusses Causal Relation Extraction Tool and looks at all the steps to extract and transform the knowledge into indexed database. Finally, we conclude with our conclusion in Section 6.

2. RELATED WORK:

Popek G.and Katarzyniak R. P [1] described an approach to extract relations. Extracting relations presented in their work based on finding relations between concepts in general needs all allowed combinations of concepts and hedges to be checked results in great complexity if done without planning. There are two main ideas for this task.

First one is to remember the sets for each property X once they are extracted. Most likely, the sets will be used multiple times and it is a waste of resources to calculate them every time. The second idea is to find dependencies between co-occurrence of relations. Using these dependencies some cases can be excluded from the search. Because of that, dependencies between co-occurrence of relations are listed and used in an algorithm for a determination and update of thesaurus to potentially reduce its complexity [20]. Another important step is to take into consideration knowledge stored in the thesauri. It should be used at least in such areas as planning.

There are many studies [2],[3], [4],[5], [6],[7], [8],[9], [10],[11], [12] that are a series of investigations that may help to shed light on the knowledge, skills, and practices that researchers and practitioners use when seeking medical literature. A thorough search of the literature is important. Incomplete literature search may result in a distorted interpretation of the body of research on a topic. Decisions that are based on incomplete information are poorly informed and may waste time, work effort, and money, especially if that information is gathered from a few familiar sources using only search terms that are familiar.

Although searching by keywords is usually highly focused, there are cases where a keywords search may produce excessive irrelevant information, particularly for words with multiple meanings. Thus, it is probably more productive to use a thesaurus that was constructed to facilitate finding material [13].

Research oriented to promote using controlled vocabularies is an extensively recognized topic in biomedical community [14]. The proliferation of biomedical terminologies and the need to use them in many health care activities, as well as in information retrieval, have increased their value as knowledge resources. Providing interoperability between different knowledge sources is also a critical issue for efficient information sharing in other communities [15].

To improve performance in detecting protein subcellular localization information the author attempted to use semantic information from the Word Net thesaurus. Furthermore, they demonstrated that syntactic and semantic information is important for the performance of this method [16,17].

Volkova [18] and other construct manually ontology for extracting entities such as: animal disease names, viruses and serotypes. They then use an automated ontology expansion approach to extract semantic relationships between concepts. Such relationships include asserted synonymy, hyponymy and causality. Bakillah and Mostafavi dealt with some problems related to the representation of fuzziness in geospatial ontologies, and fuzzy semantic mapping between fuzzy geospatial ontologies [19].

PPInterFinder—a web-based text mining tool to extract human PPIs from biomedical literature. PPInterFinder uses relation keyword co-occurrences with protein names to extract information on PPIs from MEDLINE abstracts and consists of three phases. First, it identifies the relation keyword using a parser with Tregex and a relation keyword dictionary. Next,it automatically identifies the candidate PPI pairs with a set of rules related to PPI recognition. Finally, it extracts the relations by matching the sentence with asset of 11 specific patterns based on the syntactic nature of PPIpair. PPInter Finder is capable of predicting PPIs with the accuracy of 66.05% on AIMED corpus and out performs most of the existing systems [20].

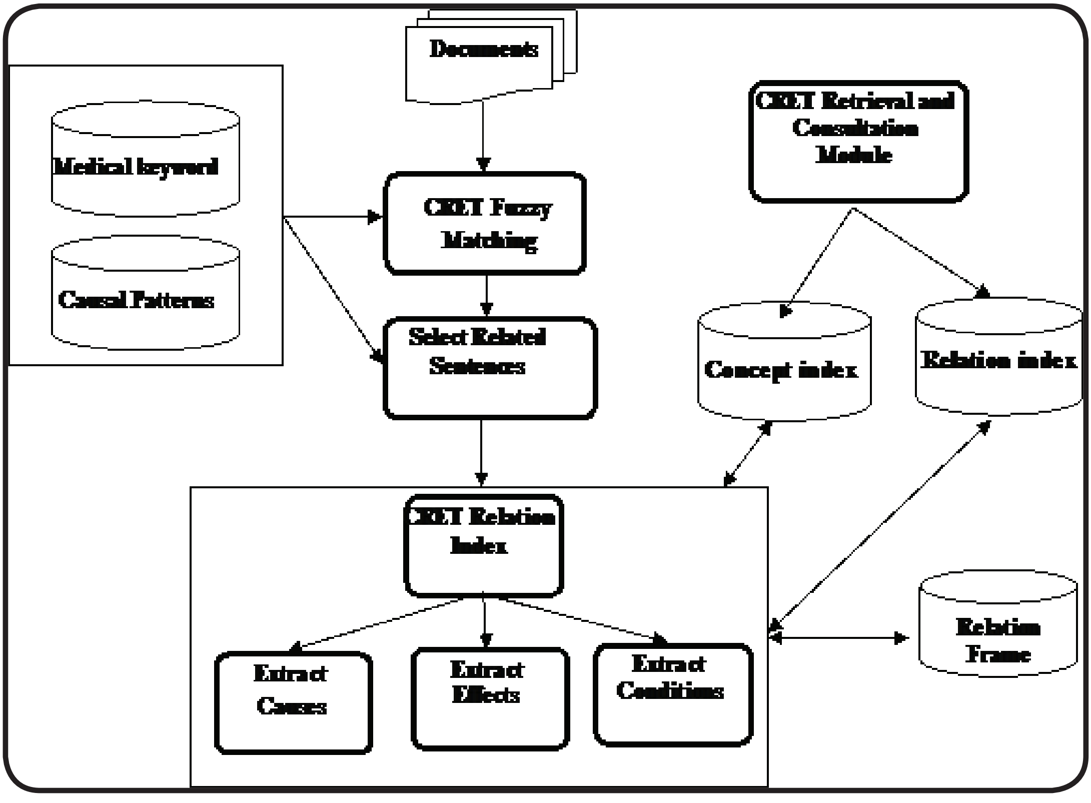
Shamsfrad introduces some lexico-syntactic and semantic patterns and templates for extracting conceptual knowledge from texts [21].

3. System Architecture:

Our system uses a relation-frame indexing to generate verb-patterns representing causal knowledge in the indexed document: each sentence is parsed for causes that are influencing certain effects within certain conditions (Figure. 1).

Figure 1:

Architecture of CRET Tool



For this purpose, a computational method has been developed to extract causal, effects and conditional sentences from texts belonging to medical domain, using them as a database to study imperfect causality and to explore the causal relationships of a given concept by means of friendly user interface. The process is divided into two major parts.

The first part, Relation Parser is an engine that generates a causal knowledge base by means of automatic detection and classification processes which are able to extract those sentences matching any of the causal patterns selected for this task.

The second part is the Relation Indexing Engine, proposes an automatic indexing mechanism which selects those sentences related to an input concept and creates an index of them, retrieving the concepts involved in the causal relationship such as the cause and effect nodes, its conditions and the type of causal relationship.

The new causal relationships and new terms generated, with nodes and relationships denote the intensity with which the causes or effects happen. This procedure should help to explore the role of causality in different areas such as medicine, biology, social sciences and engineering.

4. Documents Collection and Training:

A set of training and testing documents were collected in the domain of Medicine. The set of documents consist of abstracts and short papers in the field of medicine.

In the training phase, a medical dataset rich in medical keywords and causal relationships was used in the experiments to analyze the system developed. This database comprises of 45 documents collected from online open access medical journals. The dataset was pre-filtered and only those documents containing medical keywords likes “inflammation”, “chronics”, “diabetes”, “cancer”, “tuberculosis”, “lung”, “bronchitis”, “coronary”, “artery” etc. were added to the corpus. Each sentence from every test document was then added as a separate tuple in a sentence table. Thus, the corpus has a test set of about 1000 sentences related to medical domain.

Among those sentences, the ones expressing causation knowledge are picked out and analyzed for designing the templates (Patterns). Moreover, initial lexicons for the expected semantics based on the causation semantic templates of that particular domain are prepared manually by examining those selected sentences. The initial lexicons act as initial activations of CRET extractions module.

In our research, aiming to extract causal relationships from medical documents, we created a corpus containing 200 documents related to medicine with causal relation information. In particular, we used abstracts and short articles included on the medical domain. This corpus is useful for investigating and testing the presentation of causal relation instances and where these instances are present in these electronic documents.

After the sentences are collected from the corpus, we process these sentences to be able to easily extract the pattern or verb connecting the medical keywords of the relationship. We implemented a parser to process sentences by reducing words to their base form and assigning the patterns to each of the words in a sentence. Then, we locate the medical keywords and the input relation within the processed sentence and return the keywords between them (if it exists) and the causal relations.

5. Causal Relation Extraction Tool (CRET):

There are many relations expressed in natural language. The causation relation is an important one for human reasoning and plays an essential role in human decision making.

Our goal of Causality Relations Extraction Tool (CRET) is to develop a system able to identify sentences or passages of texts containing related information and filling slots or structured frames. These slots define entities relevant to the topic of interest and can be used for or focus on extract cause, effect and conditional information from medical texts. We improved our research by expanding the lexical approach to a semantic or conceptual level, including fuzzy relationships.

5.1 Relation Parser:

CERT uses relation patterns-based knowledge extraction approach for extracting causation knowledge from texts. In this approach, a set of generic frames is designed to focus on particular causal patterns. The patterns represent different words and sentences structures of causal relations. The patterns indicate the presence of a causal relation and which parts of the sentences represent which roles in the causal situation. Any part of the sentence that matches a particular pattern is considered to describe a causal situation, and the keywords in the sentence that match slots in the pattern are extracted and used to fill the appropriate slots in the cause-effect-condition template. In this section, we present in detail the characteristics and structures of each kind of frame and how the frames are organized to facilitate the causation knowledge extraction.

To demonstrate the usability of CERT, we study the application of CERT on medical domain. Medical knowledge is mostly expressed as causal expressions, using verb-based clauses that combine keywords or nouns, such as:

"x is reduced using y".

"x is influenced by y under conditions z" or

"x is reduced by y due to z".

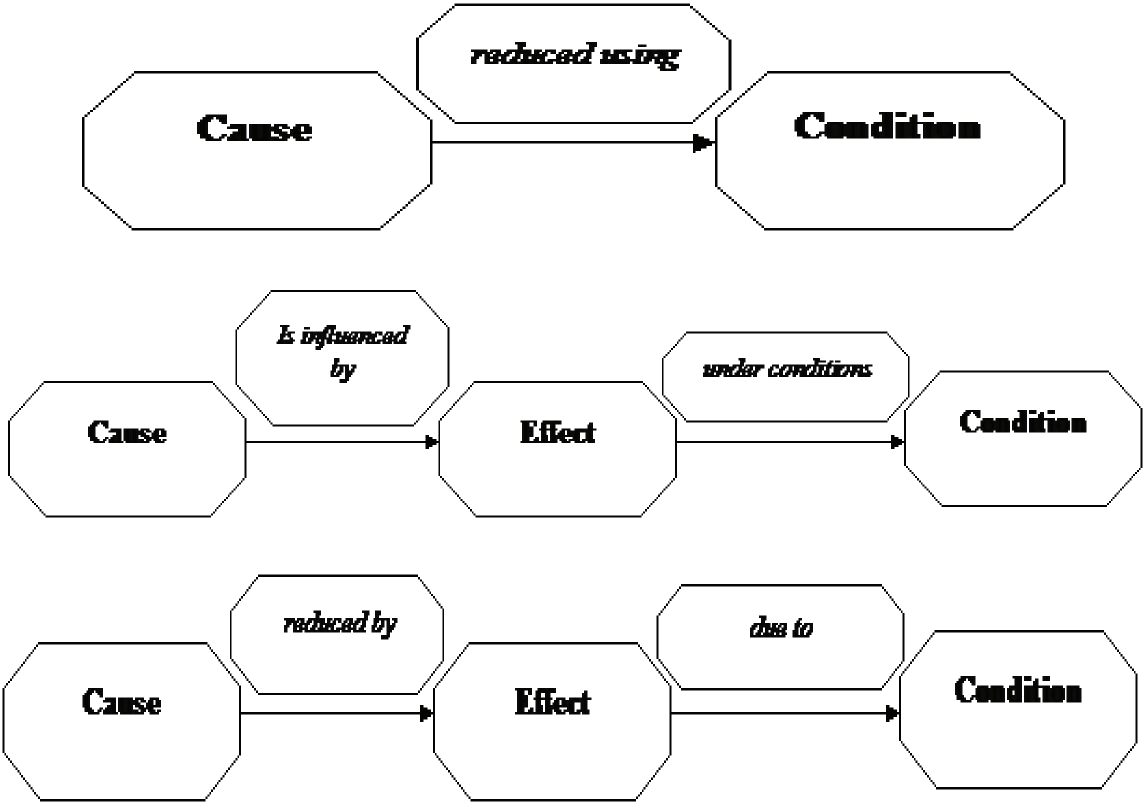
Causation relation may be regarded as one of the fundamental semantic relations. Some of the basic semantic relations, such as causation are usually expressed in structured forms. A causation relation typically has two kinds of entities, namely, a reason and a consequence. To process certain type of causal relation in texts, we represent the expected semantics of those semantic relations by semantic templates. They capture the existence of different entities or actions and their linkage. Figure 2 shows a sample of these patterns. These semantic templates capture the fact that one or more reasons cause the occurrence of a consequence.

The causation knowledge, expressed in English sentences in texts, can be categorized into different structures according to the organization of the reasons and consequences. Figure 2 shows a sample of different sentence templates in our CERT system. The first sentence template is used to model the sentence structure of simple sentences. Sentence template 1 illustrates that some cause is reduced by conditions, where as sentence template 2 illustrates that a cause is influenced by some effects under some conditions. Sentence template 3 states that some cause is reduced by effects, and those causes can come before and after the conditions.

Causation Expression in the mentioned templates refers to phrases for linking a consequence to a reason. Some examples are {as, due to, because of, because, cause, caused by, helped by}. Reasons are joined among themselves with the conjunction terms such as {and}.

Figure 2

Sample Sentence Frames



In the previous sample frames, the order of the causes and consequence is different. This shows that there are different sentence templates associated with the same causation semantic template. Here are some examples of simple sentence templates from our sample test in the medical domain (Figure 3). These four examples can be represented by the first three templates in Figure 2. The list illustrates different examples which have been automatically extracted from the sample collections of the medical documents.

Figure 3

Example of Sentences Extracted by CRET

When tear production is reduced by inflammation due to chronic Dry Eye.

Reduced alloreactive T-cell activation after alcohol intake is due to impaired monocyte accessory cell function and correlates with elevated IL-10, IL-13, and decreased IFNgamma levels.

Reduced cortical activity due to a shift in the balance between excitation and inhibition in a mouse model of Rett syndrome.

Reduced lung function due to biomass smoke exposure in young adults in rural Nepal

5.2. CRET Indexing and Consultation:

We developed Causal Relation Extraction Tool (CRET), a software indexing engine, which extracts relation patterns from the medical documents. The causal patterns are detected through a fuzzy matching process between a causal patterns and partial detected string patterns in the electronic medical documents. The process of automatic indexing of causal relations is based on the concept of verb-frames. In this approach, the frames represent language patterns extracted from natural language. The indexing process compares frames with sentences in the document and generates a table of causal relations; each relation record in this table contains 3 keywords: a cause, an effect and a condition. The relations found are stored in a relation frame index database. The database can be used to analyze links between possible causes, conditions and effects.

Most causal relations, which occur in natural language sentences, can be matched with relation frames which express arguments in relations to a causal verb (Figure 4).

Figure 4

Sample Relation Frame

frame CF04: CAUSAL_FRAME_04{<x> is influenced by <y>; in ; <z> }
for which <x> is the effect, <Y> is the cause and <z> is the context.

The arguments x, y and z of such frames are variable parameters. The CRET tool has a database of these relation-frames; this database represents causal patterns used for automatic indexing of the medical documents. A first try-out list has been manually coded by an expert in co-operation with a researcher in medicine. During iterative testing of this list; the authors of the tested documents annotated the relation-frames so that the relation-frames were adapted to a robust set.

Once a large set of medical documents have been indexed by CRET, the researcher in medicine can query the system from a causal point-of-view. ***This means that following query examples easily can be solved:***

- ♦ "show all documents about the effect of cause-x on effect-y in condition z"
- ♦ "Is there a document which discusses the effects of x"
- ♦ "give me all effects of cause x under condition y"
- ♦ etc..

6. Evaluation:

The metrics we used to analyze our test results for causal relation extraction were precision, recall and F-measure. Precision is defined as the ratio of relations correctly extracted by the system to the total number of relations it extracted, and recall is defined as the ratio of relations correctly found by the system out the total number of relations extracted by the system.

$$recall = \frac{\text{\# of correct relations given by system}}{\text{total \# of possible correct relations in text}}$$

$$precision = \frac{\text{\# of correct relations given by system}}{\text{\# of relations given by system}}$$

Precision and Recall stand in opposition to one another. As precision goes up, recall usually goes down (and vice versa). The F-measure combines the two values.

$$\mathrm{F-measure} = \frac{2 * \mathrm{precision} * \mathrm{recall}}{(\mathrm{precision} + \mathrm{recall})}$$

These measures produced a high precision and recall for the causal relation testing, as seen in table-1 and figure 5.

Table 1 –

Recall, Precision and f-measure

	With Medical Dictionary	Without Medical Dictionary
Total Relations	72	
# relation found	60	43
# of correct relations	51	23
precision	85%	53%
recall	71%	32%
f-measure	77%	40%

The evaluation results are presented in Table 2. Recall is the percentage of the slots filled by the human analysts that are correctly filled by our system. Precision is the percentage of slots filled by our system that are correct (i.e. the causal relations entered in the slot is the same as that entered by the human analysts). If the causal relation entered by the system is partially correct, it is scored as 0.77 (i.e. 77% are correct). The F-measure given in Table2 is a combination of recall and precision equally weighted.

Figure 6 and table 2 show the metrics measurement for the causality relations extraction by our system.

Figure 5

Precision, Recall and F-measure

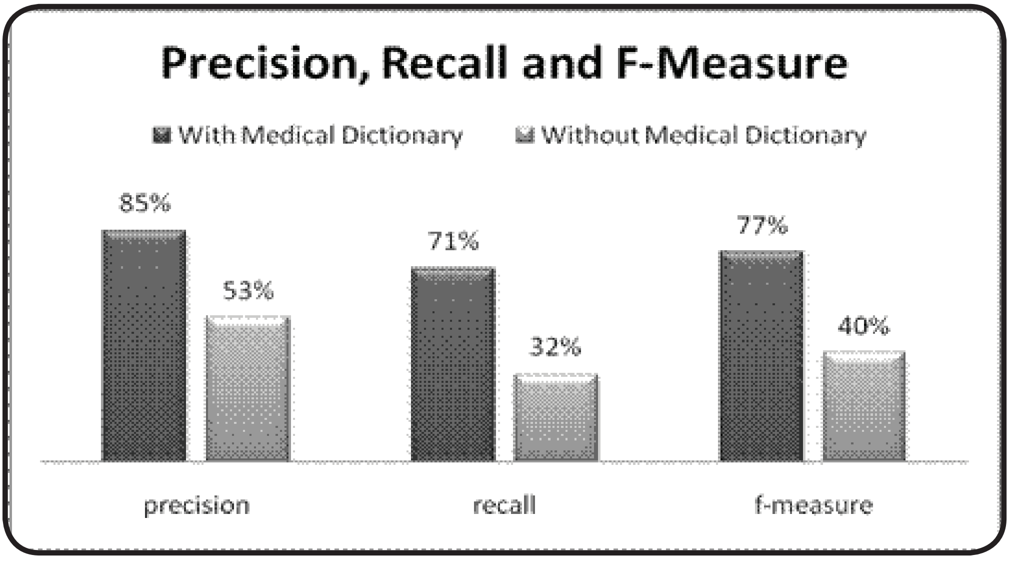


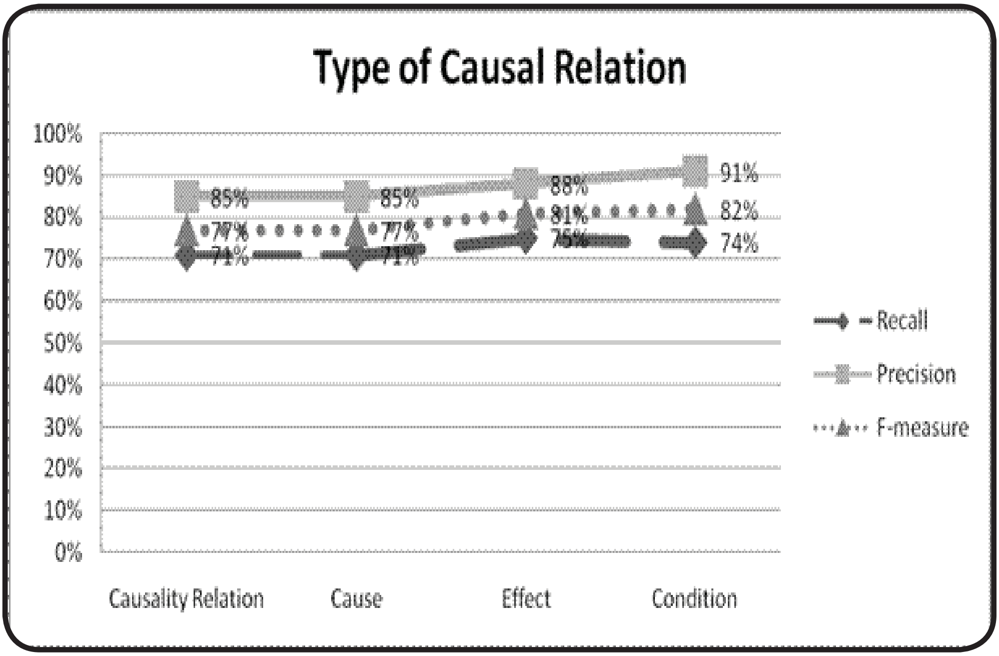
Table 2 –

Relation Extraction

Slot	Total Relations	# relation found	# of correct relations	Recall	Precision	F-measure
Causality Relation	72	60	51	71%	85%	77%
Cause	72	60	51	71%	85%	77%
Effect	60	51	45	75%	88%	81%
Condition	54	44	40	74%	91%	82%

Figure 6

Causal Relations Extraction



We can see in Table 2 that the precision is about the same as for the different type of causal frames, indicating that the current extraction patterns work equally well in with our sample collection. The lower recall indicates that new causality identifiers and extraction patterns need to be constructed.

We analyzed the sources of errors for the results of our sample collections (set of 200 test abstracts and short papers). The main sources of error can be categoried into two groups. Most of the causal relations extracted by the program as cause or effect or condition that did not identify by human expert were actually causal relations that were not medically relevant. As mentioned earlier, the manual identification of causal relations in the training test (45 documents) focused on medically relevant causal relations. In cases where the program did not correctly extract cause and effect information identified by the analysts, part of them were due to parsing complex structure of the sentences, and causality patterns have not been constructed for the causality identifier found in these sentence.

7. SUMMARY AND CONCLUSION:

We described a tool called Causal Relation Extract Tool (CRET) to identify and classify documents in a medical domain. The software generates automatically for each electronic Medical document, a set of keywords, concepts and relation frames. These sets of meta information can be used by researchers to search and rank relevant documents.

We have developed basic framework for CRET that extracts causation knowledge automatically from texts in the medical domain. The basic framework consists of two stages, namely, relation parser, and relation indexing and consultation. In the training phase, the first one is done manually by analyzing a training corpus containing relevant sentences of the domain. The remaining stages are processed automatically.

Relation-frame (CRET) indexing generates verb-patterns representing causal knowledge in the indexed document: each sentence is parsed for causes that are influencing certain effects within certain conditions. In the paper we illustrated the performance of this indexing engine through the tests on research reports in the domain of medicine.

As a training test, a set of 45 electronic documents in the domain of medicine were used. We tested the performance of our CRET approach on a set of 200 electronic documents (abstracts and short papers) in the domain of medicine. We compare the indexing results from our CRET tool with the keywords created manually by the authors. The results indicate that CRET generates 77% of the keywords and casual relations which have manually been associated by human.

References:

1. .Popek G.and Katarzyniak R. P. , Agent-based Generation of Personal Thesaurus, Proceedings of the First Asian Conference on Intelligent Information and Database Systems (ACIIDS.2009), 2009.

2. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R and Musen MA, Ontology-driven indexing of public datasets for translational bioinformatics, Centre for Biomedical Informatics, Feb. 2009.

3. Shah NH, Rubin DL, Supekar KS, Musen MA., Ontology-based annotation and query of tissue microarray data, AMIA Annu Symp Proc. 2006:709-13 .

4. Lowe, HJ; Barnett, GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA. 1994 Apr 13;271(14):1103–1108.

5. Lawrence DW. Searching the scholarly literature for injury prevention-related articles. Proceedings of the National Injury & Violence Prevention Research Conference, Oct 10 2007, Columbus, OH, USA: Nationwide Children’s Hospital and the Society for Advancement of Violence and Injury Research (SAVIR) 2007.

6. Lawrence D. W. and Laflamme L., Using online databases to find journal articles on injury prevention and safety promotion research: key journals and the databases that index them, Inj. Prev. 2008;14;91-95

7. Chang, C., Beghtol, C., Mackenzie, S., Maurice, P., Peck, S., Rogmans, W., et al.,. Thesaurus of Injury Prevention Terminology. SMARTRISK, Toronto 2003.

8. Gallagher, L.A., Thesaurus of Psychological Index Terms, vol. iv, 10th ed. American Psychological Association, Washington, DC 2005..

9. Gore, G., Searching the medical literature. Injury Prevention 9, 103–104, 2003.

10. McCray, A.T., The nature of lexical knowledge. Methods of Information in Medicine 37, 353–360. 1998

11. Lawrence D. W., Using online databases to find peer-reviewed journal articles on injury prevention and safety promotion research: a study of textword queries by SafetyLit users, Injury Prevention 2007;13:232–236.

12. Murphy L. S., Reinsch S. et all, Searching biomedical databases on complementary medicine: the use of controlled vocabulary among authors, indexers and investigators, BMC Complementary and Alternative Medicine 2003, 3:3

13. IJzereef1 L., Kamps J. and de Rijke M., Biomedical Retrieval: How Can a Thesaurus Help? LNCS 3761,pp. 1432-1448, 2005.

14. A. C. Yu, "Methods in biomedical ontology," J. Biomed. Ontol., vol. 30, no. 3, pp. 252–266, 2006.

15. Taboada M., Lal´ın R., and Mart´ınez D., An Automated Approach to Mapping External Terminologies to the UMLS, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 56, NO. 6, , page 1598-1605, JUNE 2009.

16. S. Riedel, E. Klein, Genic interaction extraction with semantic and syntactic chains, In Proceedings of ICML05 Workshop on Learning Language in Logic (LLL05), 2005

17. Kim M. Y., Detection of Protein Subcellular Localization based on a Full Syntactic Parser and Semantic Information, Proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD.2008), 2008.

18. Volkova, S.. Caragea, D.; Hsu, W. H.; Drouhard, J. and Fowles, L , Boosting, Biomedical Entity Extraction by Using Syntactic Patterns for Semantic Relation Discovery, In proceeding of 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Toronto, Canada, 2010, pp 272 - 278. DOI : 10.1109/WI-IAT.2010.152

19. Bakillah M. and Mostafavi M. A., A Fuzzy Logic Semantic Mapping Approach for Fuzzy Geospatial Ontologies, EMAPRO 2011 : The Fifth International Conference on Advances in Semantic Processing , 2011

20. Kalpana R., ,Suresh S. and Jeyakumar N., PPInterFinder—a mining tool for extracting causal relations on human proteins from literature, Database,Vol.2013,Article ID bas052, doi:10.1093/database/bas052.

21. Shamsfrad M., Lexico-syntactic and Semantic Patterns for Extracting knowledge from Persian Texts., International journal on Computer Sciences and Engineering, Vol 2, No. 06, 2010, pp 2190-2196.