

## Issue based automatic rejoinder for diseases – a confederacy based approach

Jayanthi Manicassamy\* and Dhavachelvan P.

\*Department of Computer Science, Pondicherry University, Kalapet, Pondicherry, India, jmanic2@yahoo.com

**Abstract-** Currently species that reside in the whole world is some how directly or indirectly exaggerated by diseases which is a phenomenon of medical problem. The reason behind this concern varies which profound to be hardly identified in many cases that made us in designing architecture to provide solution to this problem in this domain. In this paper a restricted issue based rejoinder (IBR) system is described which applies simple process for information retrieving from the retained document collections. Information extraction made from the databases of Unified Medical Language System (UMLS), MedicineNet, PubMed and Medline. Natural language processing (NLP) technique which is used by the system to rejoinder the issue using biomedical key terminologies. For diseases issues the extraction is done by using indexing on bases of species, signs or symptoms and treatments. In context, matching occurrence has been reformulated through alternative means for formulating better results. Apart from this, the convenience of adapting the system in the environment has also been symbolized through analysis.

Keywords: Bioinformatics, Biomedical, Information Extraction, Natural Language Processing (NLP), Pattern Matching.

### 1. Introduction

Diseases are abnormal conditions in organism that impair bodily functions associated with specific symptoms and signs. It may be caused by external factors. In human beings, diseases are referred to any condition that causes extreme pain, injuries, disabilities, disorders, syndromes, infections, etc. while in other contexts and for other purposes these may be considered distinguishable categories. When we explore answers for disease issues in biomedical we highly rely on various search engines which delivers documents that match the keywords based on the queries [1, 3, 6, 9]. In many cases, the terms disease, disorder, morbidity and illness are used interchangeably where as in some situations, specific terms are considered preferable. Most of the developed question answering (QA) systems fails to retrieve exact information from external databases. The drawback doesn't end with this, mostly the answers depends on the frequently asked question and answers which are been already set in the databases. The existing system doesn't process with dynamic Natural language processing (NLP) based on the user queries that doesn't rely with frequently asked questions. Natural language processing (NLP) question answering doesn't rely on comparing question skeleton structure with the user queries and doesn't go for restricted domain. Overlapping frequently asked questions (FAQ) with that of Natural language processing (NLP) will overcome all the drawback of the existing systems with exact information extraction apart from related information extraction with increment of precision from 30% to 40%. This is the reason behind for directing the research towards development of automatic rejoinder for disease based issues by users in natural language which have been represented in this paper. Issue based rejoinders tasks are to extract information's for the disease based issue queries from a large set of documents where the answer to the issue relies. The capability of extracting the rejoinder using natural language processing that utilizes

biomedical terminologies as keywords involving consideration of positional distances for the keyword to rejoinder the issues which vary based on the users input. The extracted information's are clustered on three bases like disease, species and treatments which depend upon the Natural language processing (NLP) query. Section 2 represents the motivation behind the development of the system. IBR architecture and working process have been represented in section 3 for which results and discussion have been represented in section 4. The last section concludes the present article with future enhancement.

### 2. Motivation

There exists several systems in various domains which involves easy and quick way to access information from electronic means access to latest information in biomedical for issues in diseases helps in better way for disease diagnosis and effective treatment for various species. Apart from this it would help to identify the stage of the disease. This involves extraction of various key information's which could be accessed through web from various databases that are most frequently updated. This motivated us in the development of the issue based rejoinder system. Apart from this the common limitation exists in considering single query. To overcome this multiple query based information extraction is involved. The system is capable of answering the issue based queries on disease by retrieving cluster based information using set of dynamic skeleton based query classification based on the study made. The skeleton identification doesn't involve in positional based where as positional involvement is adjoining multiple queries rather than in single issue rejoinder identification. The next section narrates about the IBR system architecture and working methodology.

### 3. IBR Architecture and Working Process

In this section, issue based rejoinder (IBR) for diseases in various species have been

explained. The architecture with main components represented gives a clear view about the workflow of the system. The components of the IBR system have been represented in figure1 are directly or indirectly related to one another for processing.

1. Issue Investigator
2. Document Identifier
3. Extraction Categorizer
4. Rejoinder Extractor

In “fig (1)” main components are represented by rounded edged squares and sub-components are represented by circles which are related to each other for completing diseases issue base based rejoinder process.

#### **Issue Investigator**

The disease based issues formulated by the user to the system is first investigated for whether the issue is of query or natural language. This is the primary step that will be carried out by the issue investigator which is considered to be the one of the major steps for exact classification identification for exact information extraction. The next step is to classify the issue through issue classifier which classifies the issues based on the set of skeleton structures already formulated as a base for issue identification. “Fig (2)” represents some of the set of predefined skeleton structures the issue classifier analysis the set of skeleton structures in which the user formulated issue fall which is stored temporarily through match analyzer. The issue classifier is also responsible for automatic formulation of possible skeleton structure and alternative terminologies using wordnet from the acquired predefined skeleton structure. The output of this component is utilized by the next component.

#### **Document Identifier**

This component is responsible for extraction related documents from the selective databases Unified Medical Language system (UMLS), MedicineNet, PubMed and Medline using web-services on the listed bases. The extracted documents are stored in local database temporarily for extraction the information's. The search for the related documents containing the required information is identified by using issue based skeleton structure that have been analyzed based on which related documents are extracted and stored.

#### **Extraction Categorizer**

Once the relevant documents have been retrieved, relevant passage is selected by the passage selector which is responsible for extracting the sentences from each extracted documents that could rejoinder the user issue. The extraction is done by keywords extraction and making a matching for the relevant rejoinder for the formulated user issues. The classifier is responsible for classifying the extracted information based on formulated

issues based on species, treatment, sign and symptoms which is a key for rejoinder extractor component.

#### **Rejoinder Extractor**

The rejoinder extractor is responsible for analyzing and extracting the required rejoinder from the selected sentences. The following tasks are carried out on each selected sentence of each selected documents. The foremost step is to infer logic form of the sentence to identify main verbs and then check whether the verb carry rejoinder to the issue related selective skeleton structure. Finally the biomedical keywords are identified from the sentence and made a comparison to the skeleton structure to verify the extracted rejoinder is accurate. The extracted rejoinder is considered to be unstructured which is structured by rejoinder categorizer. Apart from this the extracted rejoinder is modulated into a complete sense of sentence according to the user issue which is displayed to the user that categorized based on species, treatment, signs and symptoms which depends upon the user issue formulated to the system. Score based rejoinders display on categorized bases have been adopted for this system which is evaluated as

$$RS = \frac{\sum da + \sum ida}{\text{Total Associations}} \quad (1)$$

Here RS is rejoinder score of the specified categorization, direct association (da) = da \* 1 and indirect association (ida) = ida \* 0.5. Association represents connection with the keyword of the biomedical terminology. Direct association and indirect association are considered to be position based evaluation in this paper. Based on rejoinder score for each categorization rejoinder is displayed. Listing of categorization are also on rank bases which is evaluated as

$$C_{rs} = \sum \text{rejoinders} \quad (2)$$

Here  $C_{rs}$  is category score for rejoinder which is displayed based on the obtained score. The value of  $C_{rs}$  varies which depends on the sum of extracted rejoinder for that particular categorization. The next section discusses about the result of the system the system performance in terms of precision and recall.

## **4. Results and Discussions**

The performance of the system mainly lies in issue investigator section which has been also evaluated separately which found to have good performance. For example the user issue “causes for hemoptysis” which is classified based on the matched skeleton structure through issue investigator. For the structure “causes for x” found to be exact and “condition for x and treatment” also have related match partially but the first skeleton structure is taken for further processing in document extraction. Through document identifier related documents are extracted from UMLS, MedicineNet,

PubMed and Medline and temporally stored in local database. Extraction categorizer is responsible for extracting the relevant and related passages from the collected documents where “lung cancer is causes for hemoptysis” or “coughing up of blood from the respiratory system is a sign for lung cancer” and categorize accordingly as diseases (lung cancer) – species (Human) and sign (hemoptysis). From documents passage selected rejoinder is extracted using rejoinder extractor using biomedical keywords and extracted structured comparison is made for rejoinder found to be “lung cancer”. The extracted rejoinder is displayed based on rejoinder score and categorization rejoinder score for which sign categorization found to fall first followed by disease and species. The evaluation of the IBR system is carried on the whole system which found to have good performance which is concluded for more than fifteen evaluations carried out. Precision found to be 85% and recall found to be 87% for the evaluations carried out on the system. Their less variation in performance found while evaluating the tool. It is well known that, there exists lots of information needs related to biomedical area which vary in functionality where most of them fall online. The main novelty lies is for the system is completed developed for diseases based issue rejoinder for all species on categorization bases. The issue display doesn't get limited to single species but rejoinder is displayed for all species that have related and relevant rejoinders for the formulated user issues. The system utilization could be very beneficial for the community in the field of biomedical.

## 5. Conclusion

In the field of biomedical question answering is very restricted since they are considered to be very sensitive to answer for the questions. The system issue based rejoinder presented in this article is capable of rejoinder the formulated diseased based issue by the users. The tool is automatic the only position where the human intervention is required is formulating issue. The inputted issue is classified using skeleton structure based on the match relevant and related document is extracted from the databases UMLS, MedicineNet, PubMed and Medline. Selected passage from the document for the related rejoinder is extracted based on keywords of biomedical terminology. In context, matching occurrence has been reformulated through alternative means for formulating better results. For diseases issues the extraction display is based on score for both category and rejoinder display. From the result and evaluation made the system found to have good performance and beneficial in biomedical community.

## 6. References

- [1] Jayanthi Manicassamy and P. Dhavachelvan (2009) *International Journal of Recent Trends in Engineering (IJRTE)*, 1 (1), 550-555.
- [2] Jayanthi Manicassamy and P. Dhavachelvan (2009) *International Conference in Advances in Computing Communication & Control (ICAC3'09)*, 171-176.
- [3] Miller G.A. (1990) *International Journal of Lexicography*, 3 (4), 235-312.
- [4] O'Neil E.J., O'Neil P.E., Weikum G. (1993) *ACM SIGMOD Record*, 22 (2), 297-306.
- [5] Demner-Fushman D., Lin J. (2005), *Proceedings of the AAAI-05 workshop on Question Answering in Restricted Domains, Pennsylvania*.
- [6] Marta Sabou, Chris Wroe, Carole Goble, Gilad Mishne (2005) *Proceedings of the 14th international conference on World Wide Web, Chiba, Japan*, 190 - 198.
- [7] Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung and Wen-Lian Hsu (2006) *BMC Bioinformatics*, 7-92.
- [8] Jung-jae Kim, Piotr Pezik and Dietrich Rebholz-Schuhmann (2008) *MedEvi, Bioinformatics*, 24 (11), 1410–1412.
- [9] Arek Gladki, Pawel Siedlecki, Szymon Kaczanowski and Piotr Zielenkiewicz (2008) *Bioinformatics*, 24 (8), 1115–1117.
- [10] Benamara F. (2004) *The WEBCOOP experiment," Proc. ACL 2004 Workshop on Question Answering in Restricted Domains*, 31–38.
- [11] Niu Y. and Hirst G. (2004) *Proc. ACL 2004 Workshop on Question Answering in Restricted Domains*, 54–61.
- [12] Moldovan D., Clark C., Harabagiu S., Mairorano S. (2003) *Proceedings of the HLT-NAACL, Edmonton, Canada*, 87-93.
- [13] Ely J.W., Osheroff J.A., Ebell M. H. (2000) *Brit. Med. J.*, 321, 429-432.
- [14] Francois Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, Jean Morissette (2008) *Journal of Biomedical Informatics*, 41 (5), 706-716.
- [15] Yoshimasa Tusuroka, Junichi Tsujii and Sophia Ananiadou (2008) *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Columbus, Ohio*, 129-132.

**Fig (1): IBR System Architecture**

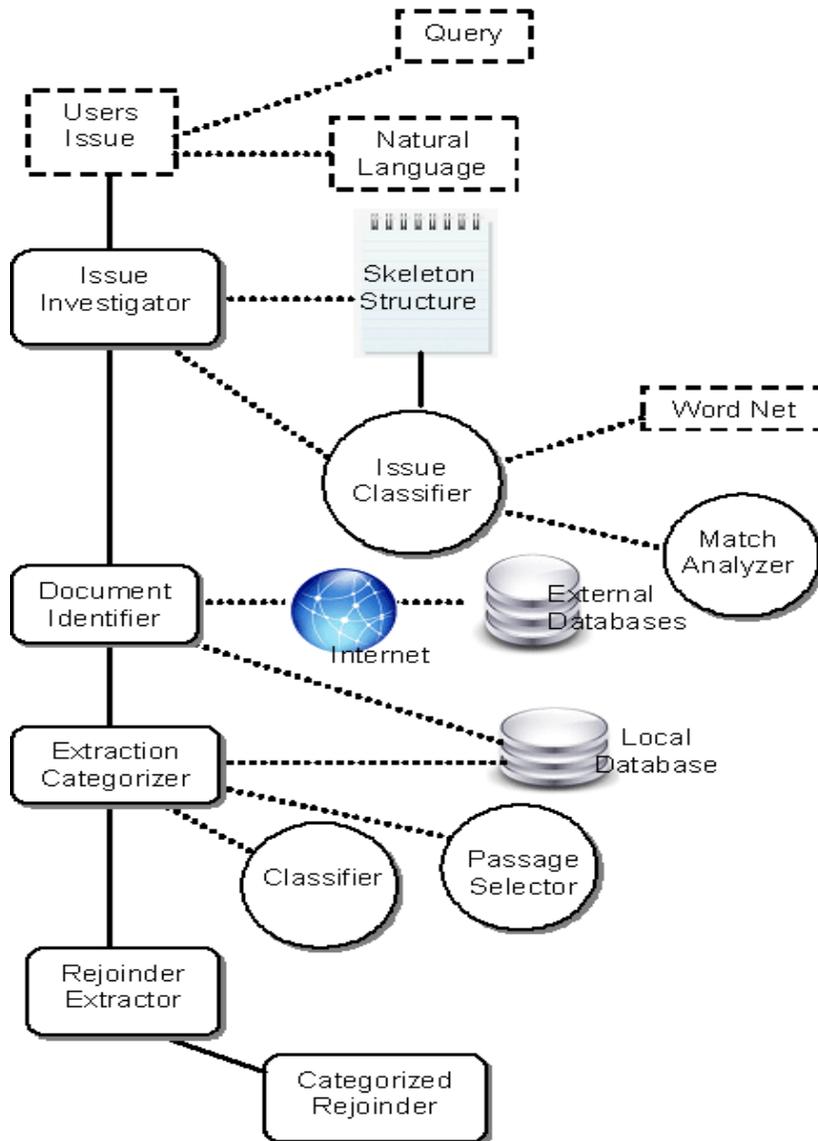


Fig 1: IBR System Architecture

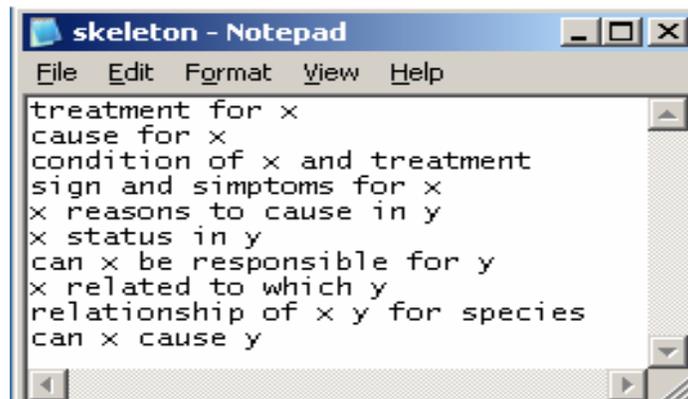


Fig 2: Predefined Skeleton Structure