iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** II **Month of publication:** February 2018

DOI: http://doi.org/10.22214/ijraset.2018.2001

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 6 Issue II, February 2018- Available at www.ijraset.com

Multiple Domain Answering by Analysing Semantic Relationship in QA Forum

J. Johnson Rajasingh, S.Balika[1]
[1]*SRM UNIVERSITY, Ramapuram, Chennai, Tamilnadu, India*

Abstract: ***Recent developments made in the web services have applied to the Information retrieval tasks. Semantic matching is a critical task for many applications in several Natural Languages processing like question answering scheme, etc. Key phrases is the subfield that contains metadata that summarizes and characterize the documents. Though, previous techniques were introduced a key phrase extraction model, still the issues like word mismatching, misidentification of the words are not yet focused. In this paper, we have proposed an efficient key phrase extraction model that efficiently retrieves the relevant data in lesser time. We have constructed machine learning models which build an index for every keyword. Firstly, the keyword is allowed for stemming process that eliminates the stop words in the sentences. Then, the stemmed words is further allowed to build into normalized words that combines with Medinet and Word net. By doing so, we have achieved faster-response time for query retrieval process of the Question Answering scheme. Experimental results have shown the efficiency of the proposed system.***

I. INTRODUCTION

Semantic matching is a critical task for many applications in natural language processing (NLP), such as information retrieval , question answering and paraphrase identification. Taking question answering as an example, given a pair of question and answer, a matching function is required to determine the matching degree between these two sentences. Recently, deep neural network based models have been applied in this area and achieved some important progresses. A lot of deep models follow the paradigm to first represent the whole sentence to a single distributed representation, and then compute similarities between the two vectors to output the matching score. In general, this paradigm is quite straightforward and easy to implement, however, the main disadvantage lies in that important local information is lost when compressing such a complicated sentence into a single vector.

Patients seeking online information about their health, connecting patients with doctors worldwide to know about their health via question and answering. Doctors able to interact with many patients about particular issue and provides instant trusted answers for complex and sophisticated problems. Previously external dictionary is used to relate medical data which was not that much sufficient enough. Here we incorporate corpus aware terminology which is used to relate the natural language medical data with medical terminology this narrow down the path between health seekers and health providers. For example: heart attack can also be said as myocardial disorder. A tri-stage framework is used to accomplish the task.

A. Noun phase extraction
B. Medical concept detection
C. Medical concept normalization

Due to loss of information global learning approach is used to complement local mining approach.

A central topic in developing intelligent search systems is to provide answers in finer-grained text units, rather than to simply rank lists of documents in response to Web queries. This can not only save the users' efforts in fulfilling their information needs, but also will improve the user experience in applications where the output bandwidth is limited, such as mobile Web search and spoken search. Significant progress has been made at answering factoid queries , such as "how many people live in Australia?", as defined in the TREC QA track.

II. RELATED WORKS

A. An Automated System for Conversion of Clinical Note into SNOMED Clinical Terminology

SNOMED-CT consists of many medical concepts and relationships. Identifies the medical concepts in the text. It is mainly used for medical data retrieval. System mainly comprises of three modules namely Augmented lexicon, term compositor and negation detector. Augmented lexicon traces the words that appears in the text and identifies the concepts that are also in the SNOMED-

CT. SNOMED-CT descriptions are then made into atomic words. UMLS Specialist lexicon performs the normalization. Normalization involves the removal of stop words. A token matching algorithm is used. It identifies the SNOMED-CT description in the text and also retrieves the related descriptions from the data structure. Matching matrix is used to identify the sequences. Negation identification is to identify the negative terms. Number of negative terms are present in clinical text . SNOMED-CT also contains numerous negation terms. For each negative term in the text there is a mapping in the SNOMED-CT. Mapping is performed based on the SNOMED-CT concept id. To detect other negative concepts a simple rule based negation identifier is used. Identifies negation terms of the form negation phrase (SNOMED CT phrase)* (SNOMED CT phrase)* negation phrase SNOMED-CT consists of many qualifier values. Qualifier modifies the medical concepts. Qualifier words are separated from augmented lexicon during concept matching. This section depicts the existing approaches carried out in the field of key phrases extractions.

B. Preliminaries

Automatic key phrase extraction systems have been evaluated on corpora from a variety of sources ranging from long scientific publications to short paper abstracts and email messages. There are at least four corpus-related factors that affect the difficulty of key phrase extraction.

1) *Length:* The difficulty of the task increases with the length of the input document as longer documents yield more candidate key phrases. For instance, each Inspect abstract has on average 10 annotator-assigned key phrases and 34 candidate key phrases . In contrast, a scientific paper typically has at least 10 key phrases and hundreds of candidate key phrases, yielding a much bigger search space. Consequently, it is harder to extract key phrases from scientific papers, technical reports, and meeting transcripts than abstracts, emails, and news articles.

D. Existing approaches

Generally, the keyphrase extraction executes in the following steps:
Extracting a list of words/phrases that serve as candidate key phrases using some heuristics.
Determining which of these candidate keyphrases are correct keyphrases using supervised or
Unsupervised approaches.

E. Selecting candidates words or phrases

As noted before, a set of phrases and words is typically extracted as candidate keyphrases using heuristic rules. These rules are designed to avoid spurious instances and keep the number of candidates to a minimum. Typical heuristics include (1) using a stop word list to remove stop words, (2) allowing words with certain parts-of-speech tags (e.g., nouns, adjectives, verbs) to be candidate keywords (3) allowing n-grams that appear in Wikipedia article titles to be candidates and (4) extracting n-grams or noun phrases that satisfy pre-defined lexico-syntactic pattern(s) .
Many of these heuristics have proven effective with their high recall in extracting gold keyphrases from various sources. However, for a long document, the resulting list of candidates can be long . Consequently, different pruning heuristics have been designed to prune candidates that are unlikely to be keyphrases.

F. Supervised Approaches

Research on supervised approaches to keyphrase extraction has focused on two issues: task reformulation and feature design.

G. Task Reformulation

Early supervised approaches to keyphrase extraction recast this task as a binary classification problem . The goal is to train a classifier on documents annotated with keyphrases to determine whether a candidate phrase is a keyphrase. Keyphrases and non-keyphrases are used to generate positive and negative examples, respectively. Different learning algorithms have been used to train this classifier, including naive Bayes.

H. Feature Selection

Structural features encode how different instances of a candidate keyphrase are located in different parts of a document. A phrase is more likely to be a keyphrase if it appears in the abstract or introduction of a paper or in the metadata section of a web page. In fact, features that encode how frequently a candidate keyphrase occurs in various sections of a scientific paper (e.g., introduction,

conclusion) and those that encode the location of a candidate keyphrase in a web page (e.g., whether it appears in the title) have been shown to be useful for the task.

1) Unsupervised Approaches
2) Graph based ranking

Intuitively, keyphrase extraction is about finding the important words and phrases from a document. Traditionally, the importance of a candidate has often been defined in terms of how related it is to other candidates in the document. Informally, a candidate is important if it is related to (1) a large number of candidates and (2) candidates that are important. Researchers have computed relatedness between candidates using co-occurrence counts and semantic relatedness and represented the relatedness information collected from a document as a graph .
This instantiation of a graph-based approach overlooks an important aspect of keyphrase extraction, however. A set of keyphrases for a document should ideally cover the main topics discussed in it, but this instantiation does not guarantee that all the main topics will be represented by the extracted key phrases. Despite this weakness, a graph-based representation of text was adopted by many approaches that propose different ways of computing the similarity between two candidates.

I. Topic based clustering
Another unsupervised approach to keyphrase extraction involves grouping the candidate keyphrases in a document into topics, such that each topic is composed of all and only those candidate keyphrases that are related to that topic. There are several motivations behind this topic-based clustering approach. First, a keyphrase should ideally be relevant to one or more main topic(s) discussed in a document. Second, the extracted keyphrases should be comprehensive in the sense that they should cover all the main topics in a document. Below we examine three representative systems that adopt this approach.
KeyCluster: The author in adopts a clustering-based approach (henceforth Key Cluster) that clusters semantically similar candidates using Wikipedia and co-occurrence-based statistics. The underlying hypothesis is that each of these clusters corresponds to a topic covered in the document, and selecting the candidates close to the centroid of each cluster as keyphrases ensures that the resulting set of keyphrases covers all the topics of the document.

III. PROPOSED WORK

This section depicts the working of the enhanced semantic architecture of the keyphrase extraction system. The thought of this proposed system arise from these issues:
To overcome from the above mentioned issues, we have proposed ranking based relevant answering systems. We have built keyphrase extraction technique that efficiently support the multiple languages. The proposed keyphrase extraction process consists of following modules:

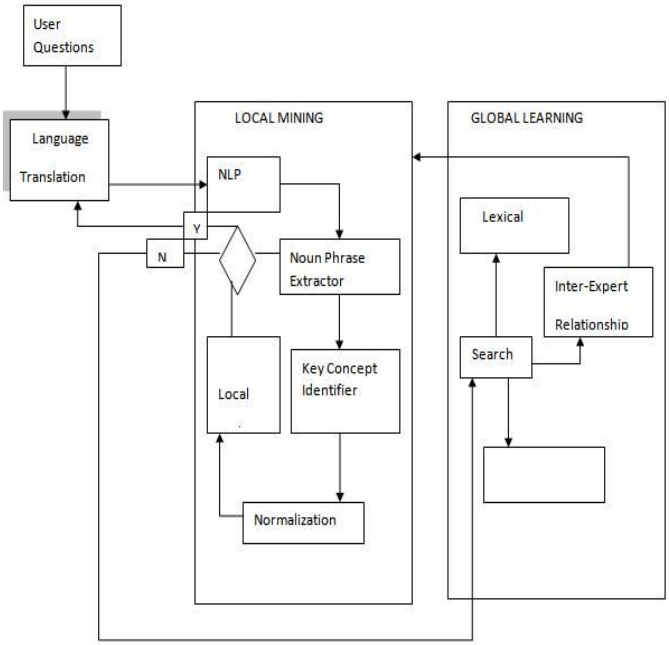
A. Q and A Application
The previous web applications posted the questions and it's answered by other users. This kind of action leads to greater redundancy and non-trusted system. In the perspective of medical practitioners, this system imposes non-trusted environment. In order to resolve this, we have built an efficient Q and A scheme that presents faster response to the answered questions and makes the user-friendly environment.

B. Key Concept Detection
The reason behind this fast-answering system is the deployment of Natural Language Processing (NLP). The objective of the NLP system is to efficiently return the answers from the relevant key terms. It specifically deals with the Parts of Speech Tagging (POST) that analyzes the phrases and nouns of the given terms. Before processing, stemming process is involved to eliminate the stopwords. This step investigates on the specific keywords from the given base words.

C. Bridging the Answers
Based on the given base words, the proper meaning will be analyzed with the help of English dictionary and medical terms. Normalization is the process executes after the completion of stemming process. A domain specific knowledge is given in the normalization process. The relevant answers are obtained from the Local Mining Database using the normalized words.

D. Machine learning And Language translation

Machine learning process operates from the use of local mining and global learning techniques. Eventually, the local mining database is updated for every given new base words. The global learning system contains a vast amount of medical related queries and terms. This will acts as backend system to retrieve the related resource to the query. An index is constructed for every keyword, so as to retrieve the words easily and at less time. If the resource is unavailable, the query will be answered later.



E. Local mining

Local Mining involves a three stage framework. First stage is the noun phrase extraction in which the noun phrase are extracted. In the second stage the medical concepts are detected using concept entropy impurity(CEI) . CEI also measures the specificity of that concept in the particular domain. Finally normalization is performed. Normalizes the medical concepts based on the authenticated vocabulary.

1)Noun Phrase Extraction: Natural Language Processing is an upcoming field in the area of text mining. As text is an unstructured source of information, to make it a suitable input to an automatic method of information extraction it is usually transformed into a structured format. Part of Speech Tagging is one of the preprocessing steps which perform semantic analysis by assigning one of the parts of speech to the given word. Part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text. Part-of-speech tagging (tagging for short) is the process of assigning a part-of speech marker to each word in an input text. Because tags are generally also applied to punctuation, tokenization is usually performed before, or as part of, the tagging process: separating commas, quotation marks, etc., from words and disambiguating end-of-sentence punctuation (period, question mark, etc.) from part-of-word punctuation (such as in abbreviations like e.g. and etc.).In noun phase extraction, it takes the speech types part into account. In this process many unwanted words are stopped because that words are uninterested meaning. To extract the noun phrases, speech tags are assigned by Stanford POS tagger to every word of medical record given by the user. Then pulls out the sequence of words that match with the fixed pattern. The input to a tagging algorithm is a string of words and a specified tag set. The output is a single best tag for each word. The noun phrases should contain zero or more adjectives or nouns, followed by an optional group of a noun and a preposition, followed all over again by zero or more adjectives or nouns, followed by a single noun. To make up a noun phrase, sequences of tags are matched in a pattern. While there are many lists of parts-of-speech, most modern language processing on English uses the 45-tag Penn Treebank tagset.

Parts-of-speech are generally represented by placing the tag after each word, delimited by a slash. For example, Take that Book. VB DT NN (Tagged using Penn Tree Bank Tag set).

2) Medical Concept Detection: Medical Concept Detection detects the medical concepts and differentiates it from other phrases. Concept Entropy Impurity is used to analyses the specificity of the medical concept. Larger CEI value indicates more relevant the concept in that domain.

3) Medical Concep: Medical concepts may not be standard terminologies so it is necessary to normalize the concepts based on a authenticated vocabulary. Consider birth control as an example it is not a standard terminology so it is necessary to map it to contraception. Authenticated vocabularies are ICD, SNOMED-CT, UMLS. SNOMED-CT provides the core general terminologies. Local mining suffer from incompleteness due to the missing key concepts. Second problem is the lower precision this is due to the irrelevant medical concepts in the records.

E. Global learning

An enhanced and novel approach of global learning is being built for enhancing the result of local coding.

1) Relationship identification: inter-terminology and inter-expert relationships are analyzed from the medical records.

2) Inter-terminology relationship: Terminologies in SNOMED-CT are arranged in hierarchies. For example, viral pneumonia is-an infectious pneumonia is-pneumonia is-a lung disease. Terminologies may also have multiple parents. For example, infectious pneumonia is also a child of infectious disease. This hierarchical representation improves the coding.

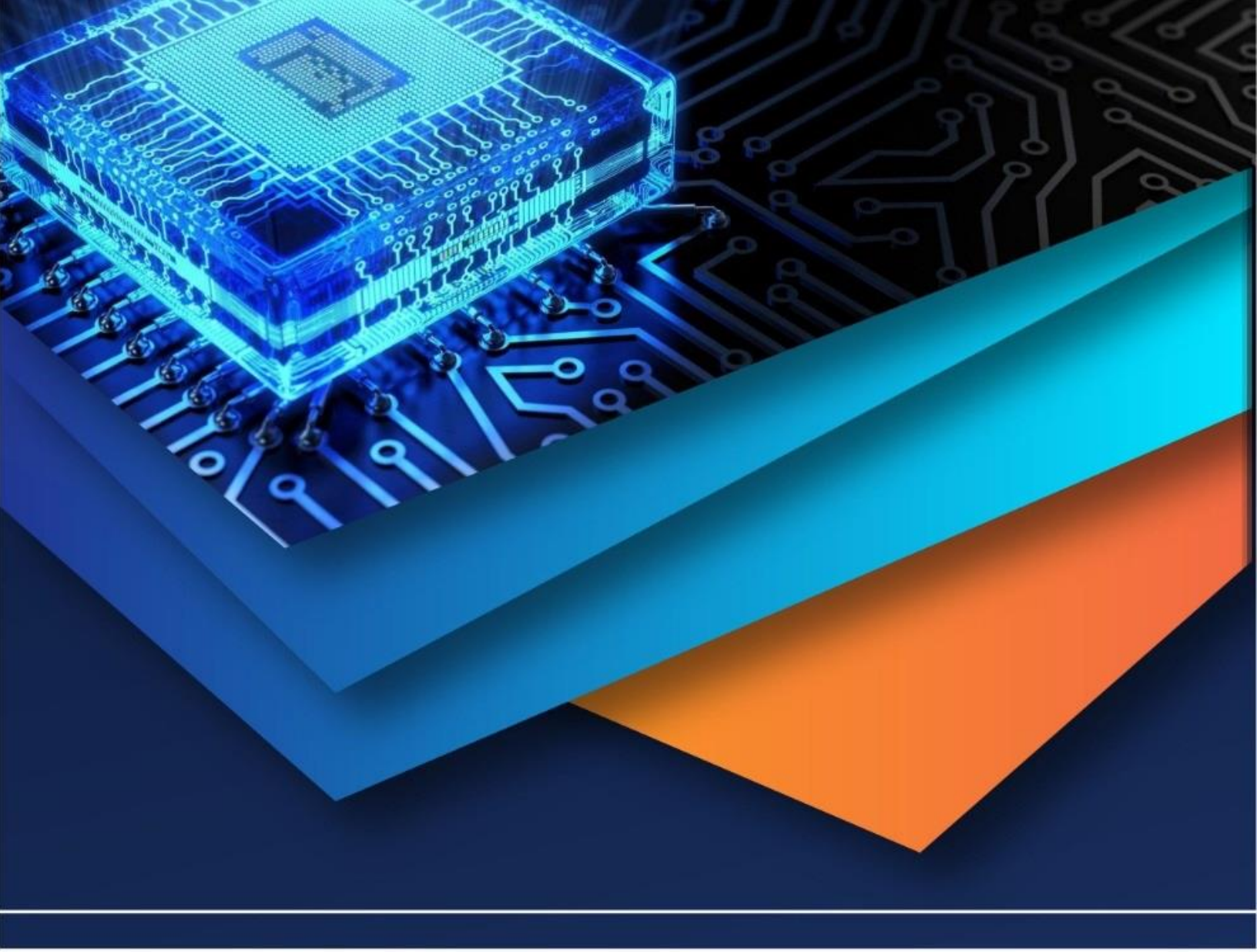
3) inter-expert relationship: analyses the historical data of experts and checks whether the experts are in the same or related area. jaccard coefficient is used to analyze the experts relationship .

III. CONCLUSION

The proposed approach consist of a combined approach within the local mining and global learning, where the corpus aware terminology is being used for making a communication between the medical support seeker and the medical care providers. The corpus terminology is having the combined approaches of local mining and global learning, where the approach of local mining undergoes within the process of stemming, noun phrase extraction, spell check, normalization and detection of medical concept. The global learning maps the query against the indexed document or keyword that is relevant to the medical records. The query is being mapped within the local database and health seekers. The output is being produced based on the patients query.

REFERENCES

[1] LiqiangNie, Yi-Liang Zhao, Mohammad Akbari, JialieShen, and Tat-Seng Chua, Bridging the Vocabulary Gap between Health Seekers and Healthcare Knowledge, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 2, FEBRUARY 2015.
[2] G. Leroy and H. Chen, Meeting medical terminology needs-the ontologyenhanced medical concept mapper, IEEE Trans. Inf.Technol. Biomed.vol. 5, no. 4, pp. 261270, Dec. 2001.
[3] Y. Yan, G. Fung, J. G. Dy, and R. Rosales, Medical coding classification by leveraging inter-code relationships, in Proc. ACMSIGKDD Int. Conf. Knowl. Discov. Data Mining, 2012, pp.193202.
[4] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute, Automating theassignment of diagnosis codes to patient encounters using example-based and machine learning techniques, J. Amer. Med. Inf.Assoc., vol. 13, no. 5, pp. 516525, 2006.
[5] C. Dozier, R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X. Guo, Fast tagging of medical terms in legal text, in Proc. Int. Conf. Artif. Intell.Law, 2007, pp. 253260.
[6] M.-Y. Kim and R. Goebel, Detection and normalization of medical terms using domain-specific term frequency and adaptive ranking, in Proc. IEEE Int. Conf. Inf. Technol. Appl. Biomed., 2010, pp. 15.
[7] S. Hina, E. Atwell, and O. Johnson, Semantic tagging of medical narratives with top level concepts from SNOMED CT healthcare data standard, Int. J. Intell. Comput. Res., vol. 2, pp. 204210, 2010.
[8] H. Stenzhorn, E. Pacheco, P. Nohama, and S. Schulz, Automatic mapping of clinical documentation to SNOMED CT, Studies Health Technol. Inform., vol. 158, pp. 228232, 2009. . Intell. Comput. Res., vol. 2, pp. 204210, 2010.
[9] Y. Chen, Z. Chenqing, and K.-Y. Su, A joint model to identify and align bilingual named entities, Comput. Linguistics, vol. 39, no. 2, pp. 229266,2013.
[10] L. Nie, M. Akbari, T. Li, and T.-S. Chua, A joint local-global approach for medical terminology assignment, in Proc. Int. ACM SIGIR Conf.,2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY