Report for MS Thesis-I

Sentence Semantic extraction for Evidence based Medicine (EBM)

Name : Muhammad Suffian

ID : FA16-MSCS-0043

Computer Science

Supervisor:

Dr. Shaukat Wasi

Mohammad Ali Jinnah University, Karachi

Table of Contents

1- Introduction

The idea of Evidence Based Medicine (EBM) caused incredible enthusiasm among wellbeing experts. As indicated by definition Evidence Based Medicine is the medication suggested by the doctors underlying the available health status of patient by formulating the question or query accordingly and then applying on the corpus of medical journals to retrieve the summaries or results related to the disease[1]. The reason for consulting the medical journals is because the medical practitioners have to get aligned with the day by day new achievements published in medical journals. The current technological advancements have revolutionized the EBM concept. This helped the doctors to opt the latest curing medications for the severe type of diseases. In spite of many hurdles, Evidence Based Medicine practice has gained reputation over recent years due to the reasons, like the improvements in patient's healthcare. Research advancements are removing the barriers in EBM and it is inferred that the boom will come with NLP techniques.

Problem Statement

Our problem is an inspiration from Sarker et al's work [2]. They discussed the problems and obstacles in evidence based medicine faced by the practitioners. They categorized the problems in five major parts.

One of those problems is related to formulate the question or query that should include all important information without ambiguity and about the information retrieval.

Natural language processing can do helpful things for the evidence based medicine. The current research in medical information retrieval has concentrated on question/query design and other facets of information retrieval to support practitioners. The sentences spoken or written by the patient are very important for the doctor and the Machine/robot to instruct/suggest/search the medication strategy from the large medical corpus or using the own skill set based on experience. The very first thing to help doctors/machines to formulate the query/strategy needs the semantic extraction or information extraction from the sentences uttered/written by the patient. Here involves the natural language processing. The correct or true Information searched or retrieved by the doctor/machine depends on the correctness of the formulation of query or the understanding developed by the doctor/machine from the sentence. The first reason is that most of the doctors and machines/robots can not formulate the correct questions because of the ambiguity in sentences due to the multiple meanings of the sentence. Second possible reason can be the less awareness of technology to doctors i.e. how to search or retrieve the information results from the corpus/journals? Now this problem of question or query formulation can be fixed using the natural language processing techniques and in this way the precision and recall of searched question or query can be increased.

2- Related Work

An approach similar in spirit to our work is discussed and modeled QRAQ [3] (Guo, Xiaoxiao, et al 2017) in which the user story as text and the challenging question is given to the agent that deduce the information from the text with existing ambiguities, and it should be able to answer the question. if the agent can not answer then firstly it learns and deduce the variables from the fact in the problem. secondly if agent can not answer the question by reasoning alone then it infers from the simulator to extract the other variables from the problem and should be relevant to question. The problem domain of this work is similar to our domain work. They used the Reinforcement Learning (RL) approach in their work and based on (RL) they presented and evaluated two memory network architectures. Our work is more towards Natural language processing and Information retrieval.

In [4] (Molla et al 2016) built a corpus for the text processing. They have taken the data set from the clinical inquiries segment of the journal dealing family practice [5]. They annotate the data using the annotation techniques like automatic extraction, manual annotation and the rephrasing text. The inquiry sentence is used as query and the retrieval text then summarized to answer. The summary of text is basically is divided into few sentence classes and the human annotation was used to classify them into according summary. They associated three evidence based answers to each question and each answer deal with separate evidence. The criteria of suggestion is based on the score of matching to the evidence. In the work of Molla et al [4] one thing can increase the accuracy of the retrieved summaries that is the removal of ambiguity from the input sentence/query.

In [6] (İlknur DÖNMEZ et al) formed a phrase-content finder system for the Turkish sentences. They have done this study by underlying the importance of subject, verb and object relation with actionable things. The phrase content relationship is also valuable because of its structural importance for sentence. They divided the sentence mainly into two parts, one the phrase and the other as content. In each sentence they separated it into 8 different phrases, then if the phrase exists the concepts are determined from the database like Word-Net. These phrase-concepts pairs like syntactic and semantic information of sentences have shown with matrix representation.

A Question/Answer system [7] (Avani et al 2017) is built focusing on the structured and annotated knowledge base. The system is divided into three parts question processing, information retrieval and the answer extraction. The question processing part is related to my study that is divided into two parts: First the question is given to python factoid question classifier [8] this determines the type of the question and also the category of answer to this question, Second the question is parsed using the Stanford dependency parser which checks the dependencies of words and POS tagging is done in parallel. In this way they determined the focus of the question. But they also highlighted the limitation of this approach that python factoid classifier does not categorized the questions in which there is a call for action. They evaluated their Question/Answer system on TREC 2004 question data set.

In [8] (kim et al 2011) build a sentence classifier that firstly identifies the key sentences and then classifies them with medical tags. Their classifier uses conditional random fields CRFs for the

learning algorithm purposes. The classifier is trained with basically four features lexical information, semantic information, structural information and sequential information. In lexical information feature they used the bag of words with bigrams and then applied POS tagging for the semantic similarity in two texts. In semantic information the metathesaurus from UMLS was used, then directly query the thesaurus with each input token. MetaMap analyzer used for sentence parsing, in this way they get the concept unique identifiers and identified the same text. The corpus was 1000 abstracts and each sentence was annotated. I highlighted only the relevant work of kim et al, their work is more towards the sentence classification retrieved from the abstracts. The features like lexical and semantic information are more related, but utilized on results after querying, the ambiguity of query and question meaning before applying on data set is not handled in their work.

In [9] (Abeed Sarker et al 2016) presented a query focused approach for text summarization to support evidence based medicine. The query specific summaries were extracted by introducing a scoring scheme in which the score was assigned to sentence on UMLS type and the category type it contains. Semantic type information improved the extractive summarization performance. They classified the questions in their corpus into medical topics using the approach [10]. For the better question associations with summaries they sets two semantic types for each question (a) important question semantic types that were identified during training and (b) important answer semantic types that is identified from human authored summaries in training. They evaluated their approach using ROUGE evaluation tool, their QSpec system outperforms previous systems working on same perspective with 96.5 % percentile rank. But the (Sarket et al) also highlighted the room for improvement that can be achieved by improving intermediate steps for the feature generation in summarization task.

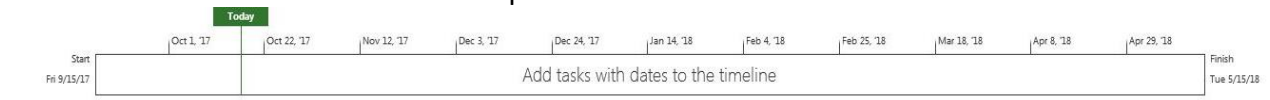
In [11](pratt et al 2000) a new approach for categorizing the search results was implemented with the name DynaCat system. In this they divided the semantics of dynamic categorization into two models (a) small query model that keeps the knowledge of the types of queries users make (b) a large domain specific terminology model, Dynacat uses UMLS for handling large terms and their synonyms. In query model the algorithm takes the types of queries and check the category of relevant query types. The limitation of query model is, it independent of disease specific terms means it generalizes the query into the specific category like categorizing in ‘treatment type’ or ‘adverse effect’ etc. This system was made for the patients and their family members with a questionnaire form to input the query data. This system was claimed better than previous ranking based and clustering based models. In this work the query or question from the patient was taken but the processing on it is not more to clear the sentence level ambiguities and to did not assigned the category on the basis of disease.

In [12] (Cao, Yonggang, et al 2011) developed an online system that is related to question answering in a complex clinical query environment, AskHERMES is a system that is in comparison with google and upToDate system for complex questions to answer with beating accuracy. Their complex question handling part is the NLP and IR problem and they have handled it with UMLS and CRFs. The system worked on vast datasets like Medline, PubMed,

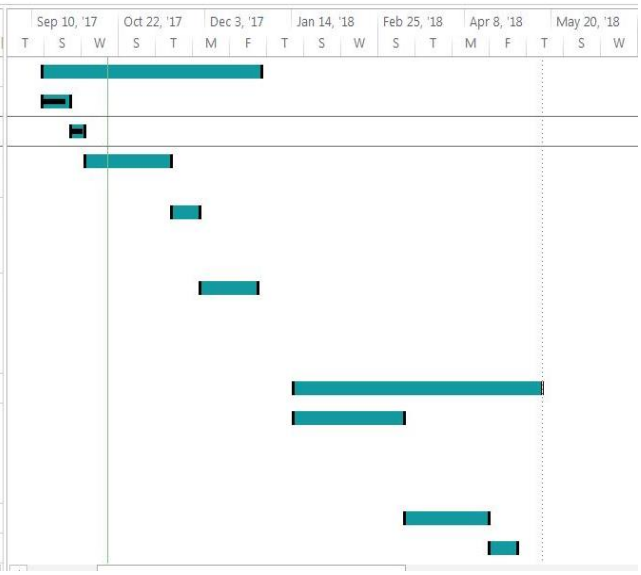
eMedicine etc. This system limit is highlighted by the Cao et al that is it does not integrate the complex clinical evidence identification part that is entered by upToDate manually.

3- Proposed Methodology

Thesis timeline with Gantt Chart The timeline for the thesis part-I and the thesis part-II is shown with the help of Gantt Chart.



Task Mode	Task Name	Duration	Start	Finish
	Thesis I	77 days	Fri 9/15/17	Sat 12/30/17
	Literaturte review	10 days	Fri 9/15/17	Thu 9/28/17
	Proposed approach	5 days	Fri 9/29/17	Thu 10/5/17
	Applying simple bi-grams on specific disease	30 days	Fri 10/6/17	Thu 11/16/17
	Evaluation on corpus of research articles on various diseases	10 days	Fri 11/17/17	Thu 11/30/17
	Proposing improvements in Language Model in terms of N(bi,tri,higher) , smoothing OR alternative approach	20 days	Fri 12/1/17	Thu 12/28/17
	Thesis II	87 days	Mon 1/15/	Tue 5/15/18
	Using improved n-Gram approach or alternative approch from Thesis-I on corpus of multiple diseaes	40 days	Mon 1/15/18	Fri 3/9/18
	Clustering of multiple diseases	30 days	Sat 3/10/1	Thu 4/19/18
	Evaluation of Improved model	10 days	Fri 4/20/18	Thu 5/3/18



Part-I:

1- Use of simple bi-grams on a corpus containing descriptions/discussions related to a specific disease

2- Evaluation on a corpus containing research articles on various diseases

3- Proposing improvements of LM in terms of N(bi or tri or higher) and smoothing OR some alternative approach for the task

Part-I Implementation

The different approaches used for the semantic extraction of sentence or patient descriptions are a) Language modeling (n-Grams) for the semantics b) keyword extraction from the patient descriptions. In figure 1 the block diagram depicts the use of different approaches, the upper part of diagram shows the –Grams approach and the lower part shows the keyword extraction approach.

1 Language Modeling:

A model that assigns the probabilities to words or sequences of words is called language model. The simplest model can be termed as **n-Gram model**. The uni-gram is each individual word having probability of occurrence, bi-gram is sequence of two words with some probability, tri-gram is sequence of three words with probability and the n-gram is sequence of n words [13].

The simple uni-grams on the plain text of a specific disease descriptions (in the form of sentences) is calculated for checking the sentence probability. The count of each word within the sentence calculated and the probability is distributed on the base of counts. The probability of the sentence is the product of each word's probability.

$$S = (w_1, w_2, w_3 ... w_i ... w_n) \quad (1)$$

The probability of sentence will be like e.i

$$P(s_{w_i}) = p(w_1 * w_2 * w_i * w_n) \quad (2)$$

In uni-grams the disadvantage is lack of context information and there is no history for co-occurrence of words or grams. According to **Markov** assumptions the history of predecessors can be stored that leads to bi-grams, tri-grams and higher grams.

$$\text{bi-gram} \sim P(w_n \mid w_{n-1}) \quad (3)$$

The bi-gram model is the probability of occurrence of a word with one word from the history. The calculation is performed by the checking the occurrence of each word with next and previous word. In this way the sentence length matters and the sequence of words also matters. If the length(vocabulary) of the sentence is n then bi-gram will lead to n^2 and the tri-gram will lead to n^3. The overall problem with the n-grams model is that if a token during the testing phase is occurred and the model didn't seen that token during training and ultimately will assign '0' probability to this token. It effects badly the sentence probability because in the case of sentence generation or checking the sentence probability, probabilities of all uni-grams/bi-grams(higher grams) are multiplied in linear fashion. So ultimately zero probability token will lead to zero probability of the sentence. To avoid this problem different smoothing techniques can be employed like add-1 smoothing discussed in section 1.2.

1.1 Evaluation of Language Modeling on patient descriptions

Evaluation of language models is done by calculating the perplexity, lower perplexity can predict the better test data by model. In case of patient descriptions (in form of the sentences) does not performed well because the new token occurring in the test set made it less accurate. An other perspective of perplexity is the branching factor, in which the future words of each word follow the given word.

$$\mathrm{PP(S)} = P(w_1 w_2 w_3 ... w_n)^{-\frac{1}{N}} \sim \sqrt{\frac{1}{P(w_1 w_2 w_3 ... w_n)}} \quad (4)$$

If we compute the perplexity with the help of bi-gram then the equation (4) becomes

$$\mathbf{PP(W)} = \sqrt[N]{\prod_{i=1}^{N} \frac{\mathbf{1}}{P(w_i \mid w_{i-1})}} \quad (5)$$

In case of patient descriptions the subject is to extract the meaning of sentence in a way that can help to make better suggestion or can help to formulate better query on the database. Also the intuition was to formulate a question by the help of n-grams for applying on the database to extract the relevant results. The n-grams didn't worked for the question creation that can later be classified into some specific disease.

1.2 Improvements through the Smoothing

The smoothing of probabilities helped the better prediction and the overall perplexity of sentence is improved. The different smoothing techniques for evaluation has been discussed by Zhai & John [14]. The Maximum likelihood method do not assign probability to the unseen word in document. The other smoothing technique is 'Jelinek-Mercer method' it is the fixed co-efficient interpolation that is used with maximum likelihood and collection model, some type of mixture model that linearly interpolated and shown in equation (6).

$$\mathrm{P}_{\mathrm{interp}}(w_i \mid w_{i-n+1}^{i-1}) = \lambda_{w_{i-n+1}^{i-1}} P_{ML}(w_i \mid w_{i-n+1}^{i-1}) + (1 - \lambda_{w_{i-n+1}^{i-1}}) P_{\mathrm{int}erp}(w_i \mid w_{i-n+2}^{i-1}) \quad (6)$$

In Laplace method the absolute discounting is preferred in the case when the unseen word in document is present that can be handled as unigram.

$$\mathrm{P}_{\mathrm{Absolutediscount}}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) - d}{c(w_{i-1})} + \lambda(w_{i-1}) P(w) \quad (7)$$

The conduct of each smoothing strategy demonstrates that, in general, the execution of longer inquiries is considerably more delicate to the decision of the smoothing parameters than that of tagged sentences.

The n-grams have been used by the M Conway et al [17] by adjoining with the other semantic features to classify the disease outbreak reports. They have shown that the combination of n-grams, bag-of-words and the semantic features can increase the accuracy of classification. n-grams solely cannot help to formulate a question or to extract the meaning from the patient descriptions/sentences. Therefore an alternative approach (section 2) is proposed that will help to initiate the query formulation or in the meaning extraction process.

2 Alternative Approach

2.1 Keyword Extraction

The keywords are the solid form of content representation in any document. Keywords help to identify the basic theme of any document. The formulation of queries in NLP and IR finds the application of keywords. Hulth have done the work of keywords extraction from the google webpages [15].

In our case the improvement in n-gram language modeling can work for the better classification or better perplexity of sentence/descriptions of diseases. But the query formulation or the prediction of the disease from the patient descriptions is needing the terms or keywords from the description. Stuart Rose, Dave Engel, Nick Cramer and Wendy Cowley wrote a chapter of text Mining book [16] in which they describe and compare the RAKE keyword extraction algorithm with different NLP based methodologies and algorithms with their specific use.

2.2.1 Keyword extraction techniques The list of keyword extraction tools :

1- TextRank (Graph based)

2- POS with supervised ML techniques

3- Chi-square measured for co-occurrences

4- Rapid automatic keyword extraction (RAKE)

5- Conditional random fields (CRM)

We have selected the RAKE algorithm for keyword extraction on the basis of its performance.

2.2 Motivation for RAKE algorithm :

The motivation for RAKE algorithm is that it is highly efficient, can operate on individual and multiple types of documents, can operate on documents which do not follow grammar rules. RAKE algorithm performs comparatively better than other enlisted algorithms/strategies in section (2.2.1) when to deal with large documents. Its extraction time is less and bug free. It is also better choice in the case when the keywords are multiple of words. It provides language and context independency.

2.3 Working of RAKE algorithm

The RAKE algorithm works on the basis of stop words, punctuations and word (phrase) delimiters. The document then parsed into candidate keywords on the basis of stop word list and delimiters.

2.3.1 Candidate keywords

The candidate keyword extraction is started by splitting the whole document into the parsed set of words in the form of array on the basis of delimiters and stop words, then the splitted array is parsed into the pairs of contiguous words which hold specific positions inside the document. These words known as candidate keywords.

2.3.2 Keyword scores

The keyword score is calculated with the help of co-occurrence graph generated for the words array. Each candidate word is placed against all the words and the co-occurrence is counted in the graph. The keyword score based on the two things mainly 1) the degree of word 2) the frequency of word. Keyword score is the ration of degree of word and the frequency of word. Now the degree of word is the co-occurrence count of candidate word with other words. The keyword score is the sum of that candidate members individual scores. If the keyword is not consists of multiple words then its ratio of degree(w) to frequency(w) is the final score. The co-occurrence graph is shown in Figure 1.

	Feeling	Fever	Yesterday	Poor	appetite
Feeling	1	1		1	
Fever		1			
Yesterday			1		
Poor	1			1	1
Appetite					1

Figure 1. Co-occurrence graph from a patient description

The calculation of keyword score is shown Figure 2.

	Feeling	Fever	Yesterday	Poor	appetite
Degree of word deg(w)	3	1	1	3	1
Frequency of word freq(w)	1	1	1	1	1
Ratio of deg(w)/freq(w)	3	1	1	3	1

Figure 2. word scores calculated from co-occurrence graph

The keyword score is the sum of word members, thus the keyword scores are:
{feeling fever : 4 , poor appetite : 4 , yesterday : 1 }

2.3.3 Extracted keywords

The keyword scores are the main role playing part to extract the keywords, so after scoring the words and finding the ratio of degree to frequency the keywords can be announced.

Results on small dataset: In our case we have applied the whole process of keyword extraction on few hand made patient descriptions and then the keywords has been announced. We used 7 patient descriptions consisting of 187 total words. The keywords extracted are shown:

Extracted keywords:

1 = [('low grade fever', 9.0), ('felling', 1.0), ('tired', 1.0)]
2 = [('stomach pain', 4.0), ('feeling hungry', 4.0), ('days', 1.0)]
3 = [('dull aching pain', 9.0), ('upper abdominal quadrant', 9.0), ('clay-colored stool presented', 9.0), ('spontaneously relieved', 4.0), ('noticed', 1.0), ('days', 1.0), ('fever', 1.0), ('jaundice', 1.0), ('passing', 1.0)]
4 = [('observed dark urine', 9.0), ('vomiting', 1.0), ('yesterday', 1.0), ('10 days', 1.0), ('age', 1.0), ('35', 0)]
5 = [('drinking tasted strange', 9.0), ('stomach flu', 3.5), ('stomach', 1.5), ('noticed', 1.0), ('beer', 1.0), ('time', 1.0), ('began', 1.0), ('feel', 1.0), ('pressing', 1.0), ('gradually', 1.0), ('lose', 1.0), ('appetite', 1.0), ('thought', 1.0), ('uh-', 1.0), ('coming', 1.0), ('weeks', 1.0)]
6 = [('breakfast', 1.0), ('suddenly', 1.0), ('fainted', 1.0), ('weeks', 1.0), ('rash', 1.0), ('body', 1.0), ('eyes', 1.0), ('skin', 1.0), ('yellow', 1.0)]
7 = [('began experiencing nausea', 9.0), ('abdominal pain', 4.0), ('darned tired', 4.0), ('double shifts', 4.0), ('young child', 4.0), ('age 26', 1.0), ('fatigue', 1.0), ('wondering', 1.0), ('working', 1.0), ('week', 1.0)]

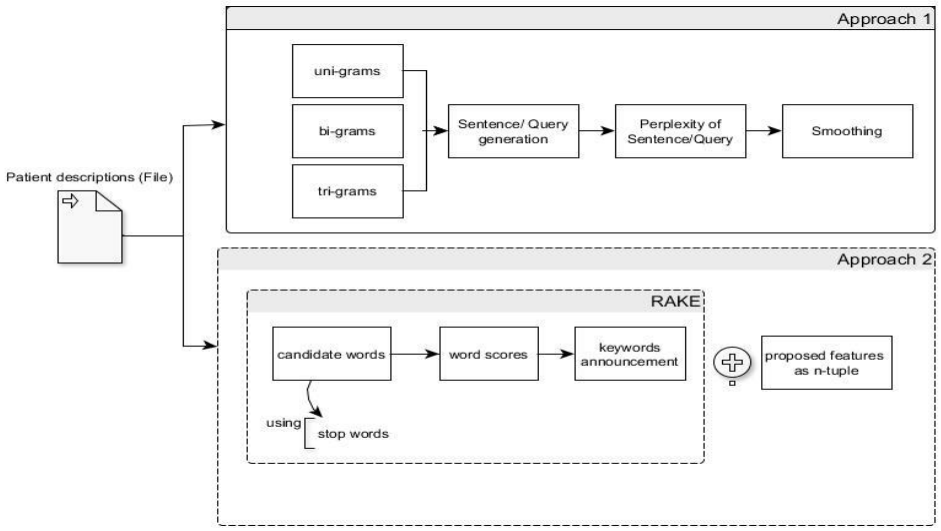


Figure (3) Block diagram for used approaches

Part-II:

1- Use of improved N-grams or the proposed alternative approach (from Thesis-I) on a corpus containing multiple diseases

2- Clustering of descriptions w.r.t diseases

3- Evaluation of improved model with clustering information

4- Finalizing the improved model

4- References

[1] Djulbegovic, Benjamin, and Gordon H. Guyatt. "Progress in evidence-based medicine: a quarter century on." *The Lancet*(2017).

[2] Sarker, Abeed, Diego Molla, and Cecile Paris. "Automated text summarisation and evidence-based medicine: A survey of two domains." *arXiv preprint arXiv:1706.08162* (2017).

[3] Guo, Xiaoxiao, et al. "Learning to Query, Reason, and Answer Questions On Ambiguous Texts." (2016).

[4] Mollá, D., Santiago-Martínez, M.E., Sarker, A. et al. Lang Resources & Evaluation (2016) 50: 705. https://doi.org/10.1007/s10579-015-9327-2

[5] http://www.jfponline.com/ [accessed on 16-10-2017]

[6] Dönmez, İlknur, and Eşref Adali. "Extracting phrase-content pairs for Turkish sentences." *Application of Information and Communication Technologies (AICT), 2015 9th International Conference on*. IEEE, 2015.

[7] Chandurkar, Avani, and Ajay Bansal. "Information Retrieval from a Structured Knowledge Base." *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on*. IEEE, 2017.

[8] Kim, Su Nam, et al. "Automatic classification of sentences to support evidence based medicine." *BMC bioinformatics* 12.2 (2011): S5.

[9] Sarker, Abeed, Diego Mollá, and Cecile Paris. "Query-oriented evidence extraction to support evidence-based medicine practice." *Journal of biomedical informatics* 59 (2016): 169-184.r.

[10] Yu, Hong, and Yong-gang Cao. "Automatically extracting information needs from ad hoc clinical questions." *AMIA annual symposium proceedings*. Vol. 2008. American Medical Informatics Association, 2008.

[11] Pratt, Wanda, and Lawrence Fagan. "The usefulness of dynamically categorizing search results." *Journal of the American Medical Informatics Association* 7.6 (2000): 605-617.

[12] Cao, Yonggang, et al. "AskHERMES: An online question answering system for complex clinical questions." *Journal of biomedical informatics* 44.2 (2011): 277-288.

[13]- Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Dan Jurafsky and James H.Martin, Draft of August 28, 2017.

[14] - Zhai, Chengxiang, and John Lafferty. "A study of smoothing methods for language models applied to ad hoc information retrieval." ACM SIGIR Forum. Vol. 51. No. 2. ACM, 2017.

[15] Hulth A 2004 Combining machine learning and natural language processing for automatic keyword extraction. Stockholm University, Faculty of Social Sciences, Department of Computer and Systems Sciences (together with KTH).

[16] - Rose, S., Engel, D., Cramer, N. and Cowley, W. (2010) Automatic Keyword Extraction from Individual Documents, in Text Mining: Applications and Theory (eds M. W. Berry and J. Kogan), John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/9780470689646.ch1

[17] Conway, Mike, et al. "Classifying disease outbreak reports using n-grams and semantic features." International journal of medical informatics 78.12 (2009): e47-e58.

5-Glossary

Appendix-A

EBM= Evidence based medicine is a technique in which the decision is taken for the health care of individuals by employing the best available evidences.

QRAQ= Query, Reason, and Answer Questions is an agent based system that learns and answer the user questions.

RL=Reinforcement Learning is technique in which the agent learns by asking the missing information.

NLP=Natural Language Processing is an Artificial Intelligence sub-area focuses on removing the ambiguity from the texts and speech.

Word-Net = word-net is a Lexical database of English Language

POS= Part of Speech is tagger that is used to specify parts of speech in a sentence.

CRF=Conditional Random Field is a classifier type statistical model that is used for pattern recognition and in machine learning.

UMLS=Unified Medical Language Systems is corpus of files and softwares that combinely works for standards of computer systems.

QSpec= Query Specific is a system that focusses on the query part.

DynaCat=Dynamical Categorisation is a system that categorises the results of queries to corpus dynamically.

JFP= The Journal of Family Practice is a place where all types of medical Q/As data sets are there.