ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Q&A On Database Using NLP

Ishant Gupta, Honey Katariya, Shaikh Mohammad Khalik, **C. D. Sukte**

BE Students, Department of Information Technology, MITCOE, Kothrud Pune, Savitribai Phule Pune University, Pune India

Professor, Department of Information Technology, MITCOE, Kothrud Pune, Savitribai Phule Pune University, Pune India

ABSTRACT: Question and Answering (QA) systems are referred as virtual assistants and are envisioned to be the next generation call center. However the accuracy of such QA systems is not as desirable and needs significant enhancement. Understanding the intent of the query is a significant contributor to an efficient system which has not been often analyzed. The current study intends to develop a QA system which can understand the query intent by using NLP based classification along with a novel scoring mechanism to extract the related information. Initially an insurance domain based simple frequently asked QA system was developed, which was then operationally enhanced into a system that can answer any type of insurance related query. The system was tested with 200 different queries and the output was cross validated by 3 experts from insurance field.

KEYWORDS-QA system; NLP algorithms; semantic similarity, user intent

I. INTRODUCTION

Computer science has always helped man in making his/her life easier. New era in human history is Information Era. With assistance of web search engines, we can get any information at our finger tips. We are just a click away from accessing a page at remote corner of the world. In addition to the brilliant research, faster computer processors and cheaper memory supported these great advancements. We have always wanted computers to act intelligent. To accomplish this task the field of Artificial Intelligence came into existence. One of the key barriers in making computers intelligent understands of Natural Language. Natural language processing which deals with understanding of languages is sub division of Artificial Intelligence. Question Answering is a classic NLP application. Task that a question answering system realizes is given a question and collection of documents, finds the exact answer for the question. It has two complementary goals: first understand the various issues in natural language understanding and representation and the second to develop natural language interface to computers [5].Natural language processing is becoming one of the most active areas in Human-computer Interaction. It is a branch of AI which includes Information Retrieval, Machine Translation and Language Analysis. The goal of NLP is to enable communication between people and computers without resorting to memorization of complex commands and procedures. In other words, NLP is techniques which can make the computer understand the languages naturally used by humans. While natural language may be the easiest symbol system for people to learn and use, it has proved to be the hardest for a computer to master [3]. Despite the challenges, natural language processing is widely regarded as a promising and critical important endeavor in the field of computer research. The general goal for most computational linguists is to instill the computer with the ability to understand and generate natural language so that eventually people can address their computers through text a though they were addressing another person.[7] The applications that will be possible when NLP capabilities are fully realized are impressive computers would be able to process natural language, translating languages accurately and in real time, or extracting and summarizing information from a variety of data sources, depending on the users' requests.The Intelligent Query Interface will participate vital role in computer interaction. It is a part of Artificial Intelligence which has information retrieval, Machine translation and Analysis [1]. The main aim of the Intelligent Query Interface (IQI) is to facilitate the novice user to interact Database by avoiding the complex command and function. This Intelligent Query Interface will make the people easy to learn and use the computer as well. [2]. This will make the user to enter the text message as they would pass to the person. The interactive with

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

computer is very essential and also more effective. Nowadays computerization is implemented in almost all the fields. Particularly in Medical Field if the Doctor wants to interact with Database, he should
Knowcomplex command as well as procedure. But this Intelligent Query processing made everyone to access the Database easily [1].

II. LITERATURE SURVEY

In this section we will be discussing the system in detail: The system we proposed in this paper converts the natural language query into SQL (Structured Query Language) which is a database programming language. We will perform the following steps for the transformation of query from natural language to database query (SQL) sequentially aslisted in the following points:Children learn language by discovering patterns and templates. We learn how to express plural or singular and how to match those forms in verbs and nouns. We learn how to put together a sentence, a question, or a command. Natural Language Processing assumes that if we can define those patterns and describe them to a computer then we can teach a machine something of how we speak and understand each other. Much of this work is based on research in linguistics and cognitive science. [1]

NLP research pursues the elusive question of how we understand the meaning of a sentence or a document. What are the clues we use to understand who did what to whom, or when something happened, or what is fact and what is supposition or prediction? While words--nouns, verbs, adjectives and adverbs--are the building blocks of meaning, it is their relationship to each other within the structure of a sentence, within a document, and within the context of what we already know about the world, that conveys the true meaning of a text.People extract meaning from text or spoken language on at least seven levels. In order to understand Natural Language Processing, it is important to be able to distinguish among these, since not all "NLP" systems use every level [2].

The morpheme is a linguistics term for the smallest piece of a word to carry meaning. Examples are word stems like child (the stem for childlike, childish, children) or prefixes and suffixes like un-, or –ation, or –s. Many new search engines are able to determine word stems on a rudimentary level. This usually means that they can automatically offer you both the singular and plural form of a word, which is a nice feature. Automatic use of morphology without accompanying understanding, however, can also return some pretty funny documents. One example that I stumbled over in PLS in its early days was that it would stem a word like “communication” and return community, commune, communication, or communism [3].

When we parsed a sentence in grade school, we were identifying the role that each word played in it, and that word's relationships to the other words in the sentence. Position can determine whether a word is the subject or the object of an action. Fred hit John and John hit Fred both use the same words, but the meaning is quite different, particularly to the person who is the object of the verb hit.[4]

NLP systems, in their fullest implementation, make elegant use of this kind of structural information. They may store a representation of either of these sentences, which retains the fact that Fred hit John or vice versa. The may also store not only the fact that a word is a verb, but the kind of verb that it is. They characterize dozens of different kinds of relationships, such as AGENT, POSSESSOR OF, or IS A. When this additional information is stored, it makes it possible to ask "Who left the golden apple for Atlanta?" without retrieving reams of information about apples in the city of Atlanta.[5]

The semantic level examines words for their dictionary meaning, but also for the meaning they derive from the context of the sentence. Semantics recognizes that most words have more than one meaning but that we can identify the appropriate one by looking at the rest of the sentence. The charm of the Amelia Bedelia books by Peggy Parish comes mostly from Amelia Bedelia confusing the senses of a word (draw the drapes, dress the chicken, dust the furniture, put out the lights).Let's look at our Lewis Carroll example again. The Jabberwocky makes sense at the syntactic level. It has some major problems at the semantic level. At the semantic level, we can finally ask what do brillig, slithy and toves mean. We found quite a bit of meaning in that sentence on the syntactic level, but it fails completely at

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

the semantic level.Some meaning is entirely dependent on context. It is context that determines which sense to assign to the verb to draw in "He drew his snickersnee" (The Mikado, Gilbert and Sullivan). Even if we don't know what a snickersnee is, the fact that the rest of the song is about an execution determines that he is talking about drawing a sword of some sort, not an artistic activity.[6]

This level examines the structure of different kinds of text and uses document structure to extract additional meaning. For instance, a newspaper article typically reports the most important facts--who, what, when, where, how--at the beginning, usually in the first paragraph. It makes predictions about the impact of the events towards the end. In contrast, a mystery novel never tells you who did it and how it was done until the end. A technical document starts with an abstract, describing the entire contents of the document in a single paragraph and then enlarging on all these points in the body of the work. [7]

Because each of these levels of language understanding follows definable patterns or templates, it is possible to inject some language understanding into a computer system by using those definitions. The higher the level, the more difficult this becomes, however. One of the greatest challenges for NLP systems is to distill a sentence or document down to an absolutely precise, unambiguous representation of its contents. These formal representations must be unambiguous and contain not only the words and meaning in a sentence, but the structure of that sentence as well. Computers, after all, are not human, and they will do only what they are told to do. [8]

III. ALGORITHM USED

1. Scan the given Input string from left to right
2. While Scanning spell check will be processed simultaneously
3. Break the sentence into several tokens by eliminating the delimiters by which the meanings of symbols will be defined by the grammar
4. Make interrelated Token list
5. Assign Rank value for the Tokens
6. create parsed syntax tree
7. convert the leaves of tree to particular SQL
8. Dependency Structure used to accumulate information.
9. Split the Query and extract patterns
10. Look for sentence connectors
11. Split Query on the basis of Connectors.
12. Specify the conditions in Query by using the Criteria
13. Find attribute and values
14. Map values with respective tables.
15. Check for validity of the data
16. Translate into SQL Query

IV. FLOW OF THE SYSTEM

The inputs to the flowchart have both the queries i.e. Normal query and also the human language query. The flowchart is shown in Fig.2.4.When the input enters, the spell checker will be processed simultaneously. After getting the input it undergoes check for IQI, if it is false then it is normal query, so it will be passed directly to query process. If it is true then the IQI is undergone the process of lexical, syntax and semantic analysis. The error handling is done accordingly for all process. After the analysis process it will send to query process, there the query will be processed and then the results will be displayed [5].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

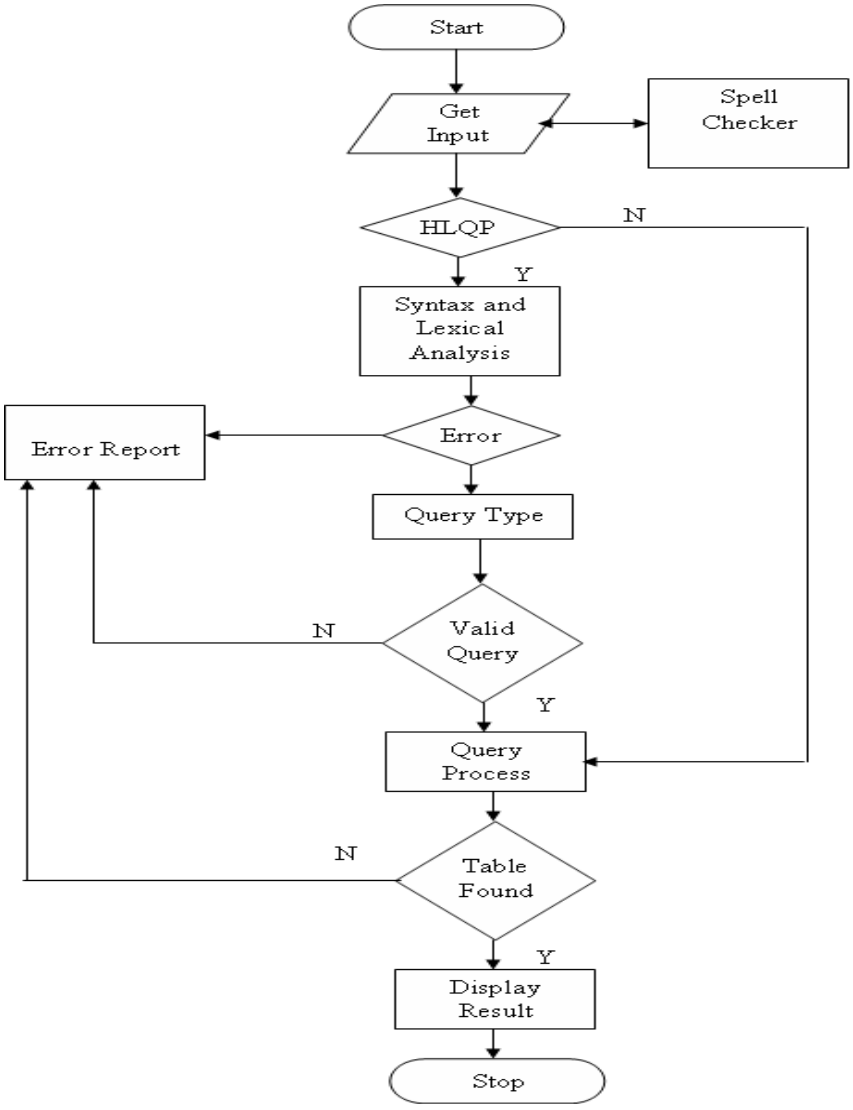

Fig No 01 Flow of the system

V. PROPOSED SYSTEM ARCHITECTURE

This paper work uses NLP for interfacing with the Database using natural language. In this work only English language is used as a mean for providing inputs. In this system we consider a database ORACLE and a default table is used which is properly normalized. A system is developed that eliminates the problem of normal user to interact with database with rigid language SQL. The users are able to access information's by issuing query in simple English language. The system is developed in JAVA language. The methodology adopted for preparing this software can be classified through following steps [10].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

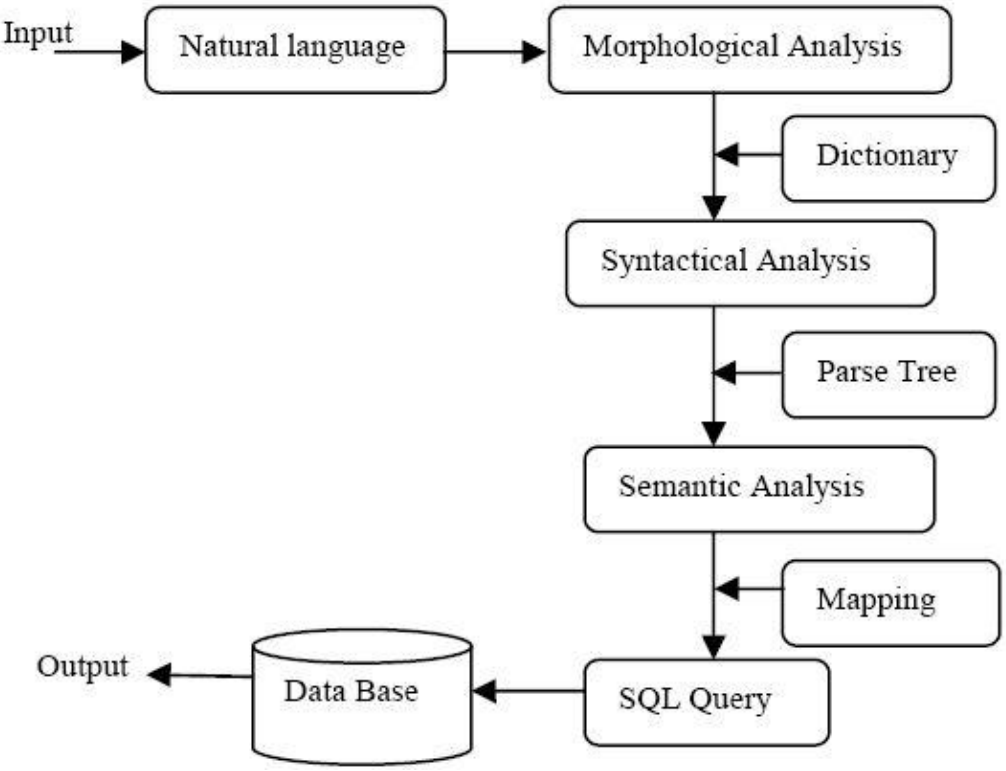

Fig No 02 Proposed System Architecture

1. Morphological Analysis: Individual words are analyzed into their components and non-wordtokens such as punctuation are separated from the words.
2. Syntactic Analysis: Linear sequences of words are transformed into structures that show how the words relate to each other.
3. Semantic Analysis: The structures created by the syntactic analyzer are assigned meanings.
4. Discourse integration: The meaning of an individual sentence may depend on the sentences that precede it and may influence the meanings of the sentences that follow it.
5. Pragmatic Analysis: The structure representing what was said is reinterpreted to determine what was actually meant.

Then the Morphological analysis are identified each world like what, is, the, Saravjeet's, RollNumber. Individual words are analyzed into their components, and it separated noun andadjective in the sentences. Morphological analysis must pull apart the word "Saravjeet's" intothe proper noun "Saravjeet" and the possessive suffix "'s". A limited data dictionary is also used to store all related words about the system. After this, Syntactic rules checks the grammatical mistakes of a sentence and Semantic analysis must map individual words into appropriate objects in the knowledge base or database and the meanings of the individual words combine with each other and find out the meaning of simple English query. Example: Meaning of query: RollNumberof student with name Saravjeet.

VI. EXPERIMENTAL SET UP

A set of 200 queries (100 generic and 100 specific) and answers for them were framed from a randomly selected document. For each topic one query was selected and the corresponding answer were either extracted from the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

document directly or manually written by understanding the concept. This manually documented set was taken as reference for evaluating our algorithm along with cross validation performed by 3 insurance domain experts. Every query from the set of 200 was provided as input to the system. Metrics such as precision, recall, F-measure, accuracy andrejection were computed to evaluate the system efficiency. Fmeasurewas further grouped into two namely F1 and f1 as was done by a previous study on word pair similarity [2]. F1 is defined as uniform harmonic mean of precision and recall, whereas f1 is defined as uniform harmonic mean of rejection and recall. System was expected to analyze and apply aforementioned logic and extract the possible short answer that is related to the user query. The responses were provided to 3 experts in insurance domain for validating the system accuracy.

An example computation of the metrics for a query, which was expected to retrieve 40 sentences as per the three experts, is explained below. Out of these 40 sentences 15 were bound to be related and the remaining unrelated. However the current QA system extracted only 20 sentences for that query. Subsequently all the extracted sentences were subjected to expert verification. It was verified that 9 sentences were related to the query and remaining were not related. Based on the standard definitions, all the related metric were found to be: Recall=0.6, Precision=0.45, Rejection=0.44,Accuracy=0.225, F1=0.514 and f1=0.507. Mean metric scores forall the queries that were put forth to evaluate the system is shown (see Table 1).

TABLE I. TABLE TYPE STYLES

Metrics	Query Type	
	Generic	*Specific*
Recall	0.6±0.025	0.8±0.02
Precision	0.45±0.04	0.667±0.08
F1	0.514±0.03	0.727±0.023
Accuracy	0.225±0.025	0.5±0.06
Rejection	0.44±0.08	0.667±0.02
f1	0.507±0.06	0.727±0.02

In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).A system with high recalls but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will return many results, with all results labeled correctly.

Precision (P) is defined as the number of true positives (T_p) over the number of true positives plus the number of false positives (F_p).

$$P = \frac{T_p}{T_p + F_p}$$

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Recall (R) is defined as the number of true positives (T_p) over the number of true positives plus the number of false negatives (F_n).

$$R = \frac{T_p}{T_p + F_n}$$

These quantities are also related to the (F_1) score, which is defined as the harmonic mean of precision and recall.

$$F1 = 2\frac{P \times R}{P + R}$$

It is important to note that the precision may not decrease with recall. The definition of precision ($\frac{T_p}{T_p + F_p}$) shows that lowering the threshold of a classifier may increase the denominator, by increasing the number of results returned. If the threshold was previously set too high, the new results may all be true positives, which will increase precision. If the previous threshold was about right or too low, further lowering the threshold will introduce false positives, decreasing precision.

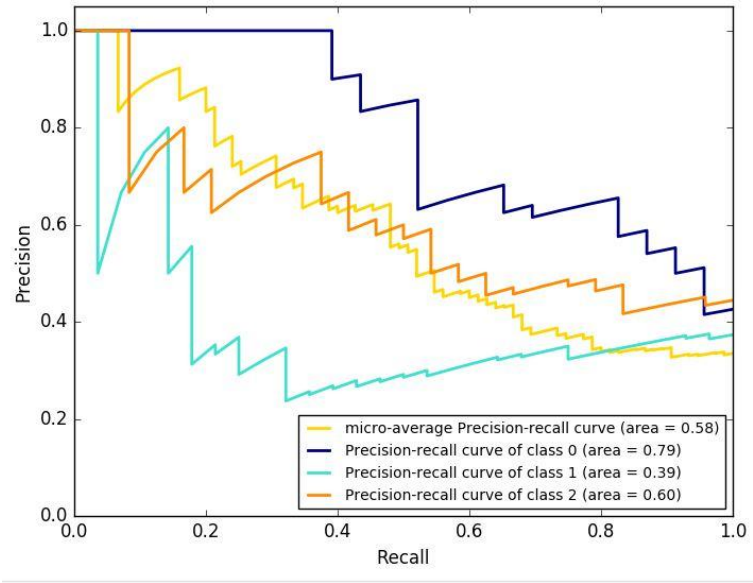

Fig No 03 Precision and Recall

VII.CONCLUSION

From the consideration of all the above points we conclude thathowever the accuracy of such QA systems is not as desirable and needs significant enhancement. Understanding the intent of the query is a significant contributor to an efficient system which has not been often analyzed. The current study intends to develop a QA system which can understand the query intent by using NLP based classification along with a novel scoring mechanism to extract the related information. Initially an insurance domain based simple frequently asked QA system was developed, which was then operationally enhanced into a system that can answer any type of insurance related query. The system was tested with 200 different queries and the output was cross validated by 3 experts from insurance field.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

REFERENCES

[1]Huangi,.GuiangZangi, phillip C-y Sheu A Natural Language database interface based on probabilistic context free grammar". IEEE international workshop on Semantic Computing and systems 2008
[2] Akama, S. (Ed) Logic.1onguage and computation. Kulwer Academic: publishers. pp. 7-11, 1997.
[3] ELF Software CO. Natural-Language Database interfaces from ELF Software Co. cited november 1999. Available from internet
[4] Hendrix. G.G, Sacerdoti, E.D,sagalowicz. D. Slocum. J. '"Developing a natural Language interface to complex data in ACM Transaction on database system. 3(2). pp. 105- 147,1978.
[5] http://www.cnlp.org/publications/03nlp.lis.encyclopedia.pdf
[6] http://www.enggjournals.com/ijcse/doc/IJCSE10-02-02-20.pdf
[7] Joseph, S.W., Aleliunas, R. "A knowledge-based subsystem for a natural language interface to a database that predicts and explains query failures", in IEEE CH, pp. 80-87, 1991.
[8] Mitrovic, A. A knowledge-based teaching system for SQL,University of Canterbury, 1998. Moore, J.D. "Discourse generation for instructional applications: making computer tutors more like humans", in Proceedings AI-ED, pp.36-42,1995.
[9] Suh, K.S., Perkins, W.C., "The effects of a system echo in a restricted natural language database interface for novice users", in IEEE System sciences, 4, pp. 594-599, 1994.
[10] Whenhua, W., Dilts, D.M. "Integrating diverse CIM data bases: the role of natural language interface", in IEEE Transactions on systems, man, and cybernetics, 22(6), pp. 1331-1347, 1992.
[11] [9] Dan Klein, Christopher D. Manning: Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency. ACL 2004: 478-485
[12] Anxerre P and Inder R. MASQUE Modular Answering System for Queries inEnglish - User Manual. AI Applications Institute, University of Edinberg, tech.rep.
[13] Broder, A. Z., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V., & Zhang, T. (2007, July). Robust classification of rare queries using web knowledge. In Proceedings of the 30th annual
international ACM SIGIR conference on Research and development in information retrieval (pp. 231-238). ACM.[14] Shen, D., Sun, J. T., Yang, Q., & Chen, Z. (2006, August). Building bridges for web query classification. In Proceedings of the 29thannual international ACM SIGIR conference on Research and development in information retrieval (pp. 131-138). ACM.
[15] Shen, D., Pan, R., Sun, J. T., Pan, J. J., Wu, K., Yin, J., & Yang, Q. (2005). Q 2 C@ UST: our winning solution to query classification in KDDCUP 2005. ACM SIGKDD Explorations Newsletter, 7(2), 100- 110.
[16] Cao, YongGang, et al. "AskHERMES: An online question answering system for complex clinical questions." Journal of biomedical informatics 44.2 (2011): 277-288.
[17] Athenikos, Sofia J., and Hyoil Han. "Biomedical question answering: A survey." Computer methods and programs in biomedicine 99.1 (2010): 1-24.
[18] Janzen, Sabine, and WolfgangMaass. "Ontology-based natural language processing for in-store shopping situations." Semantic Computing, 2009.ICSC'09.IEEE International Conference on.IEEE, 2009.
[19] "The Stanford NLP (Natural Language Processing) Group." The Stanford NLP (Natural Language Processing) Group. N.p., n.d. Web. 06 Jan. 2015.
[20] "Apache OpenNLP Developer Documentation." Apache OpenNLPDeveloper Documentation. N.p., 06 Jan. 2015. Web. 06 Jan. 2015.