International Journal of Computer Applications (0975 – 8887)
Volume 65– No.12, March 2013

Effective Question Answering Techniques and their Evaluation Metrics

Jaspreet Kaur
ME, Computer Science and Engineering
University Institute of Engineering & Technology
Panjab University, Chandigarh

Vishal Gupta
Assistant professor, Computer Science and Engineering Department
University Institute of Engineering & Technology
Panjab University, Chandigarh

ABSTRACT

Question Answering (QA) is a focused way of information retrieval. Question Answering system tries to get back the accurate answers to questions posed in natural language provided a set of documents. Basically question answering system (QA) has three elements i.e. question classification, information retrieval (IR), and answer extraction. These elements play a major role in Question Answering. In Question classification, the questions are classified depending upon the type of its entity. Information retrieval component is used to determine success by retrieving relevant answer for different questions posted by the intelligent question answering system. Answer extraction module is growing topics in the QA in which ranking and validating a candidate's answer is the major job. This paper offers a concise discussion regarding different Question Answering types. In addition we describe different evaluation metrics used to evaluate the performance of different question answering systems. We also discuss the recent question answering systems developed and their corresponding techniques.

General Terms
Architecture, Reliability Assessment, Evaluation.

Keywords
Natural Language Processing, Question Answering System, Evaluation Metrics, Information Retrieval.

1. INTRODUCTION
Recognizing user questions in natural languages involves Natural Language Processing (NLP).Being an important field of research, NLP plays a big role in the Information and Communications Technology (ICT) and Question Answering (QA) systems. Natural language processing (NLP) is the automated approach to scrutinizing text based on both a set of technologies and a set of theories. Rather than the keyword based retrieval methods, it has become significant to be able to ask queries and get answers, using natural language (NL) expressions [1]. The QA system can better fulfill the needs of users as they provide with a perfect, faster, suitable and successful way of giving answers to user questions.

There is a speedy development in information technology, and people's desire for faster and accurate output promoted the growth of Question Answering Technology. A Question Answering System is one of the major key areas for the researchers in recent years. Lot of work is already done in this area [2].The purpose of this system is to examine the user query intelligently and provide the most appropriate answer to the user. The main areas under Question Answering system are the Query Optimization and the Query Accuracy. The Optimization basically deals with the response time to find the query result and Accuracy is the degree of user satisfactory ratio. Question Answering systems are being defined by different researchers in different frameworks like Network based QA System, Web Based System, Mobile Based System etc. With each system the main objective is to get the most accurate output with least consumption of resources and the time. In 1999, the first TREC question answering (QA) track recognized as a objective the retrieval of short answers instead of documents under the assumption "that users would usually wish to be given the answer instead to find the answer at their own in a document" [3].QA and other focused-retrieval systems [4] seek to remove much of a user's expenses in finding information. At the front end, the user of a QA system needs to enter a natural language question instead of a keyword query. The ideal QA system obtains, extracts, and composes a brief answer and saves users time that they would have to spend on these tasks if they used a document retrieval system[5].

In this paper, we will be discussing the basic structure of Question Answering System and its classifications. Further we will be discussing different types of Question Answering systems developed and different evaluation metrics used to evaluate the performance of Question Answering Systems.

2. FUNDAMENTALS OF QUESTION ANSWERING SYSTEM
Question Answering (QA) System is a programmed system having the capability of answering natural language questions in a human like manner with quick and perfect result. In 60's, since the early days of artificial intelligence, researchers have been captivated with answering natural language questions. However, the obscurity of natural language processing (NLP) had restricted the scope of Question Answering to domain-specific expert systems. Question Answering has been considered in NLP since 1970's with the systems like BASEBALL [6], which provides answers to questions about the American Baseball League and LUNAR [7], which allowed geologists to ask questions concerning moon rocks. In recent years, web growth, progress in information technology and the explosive demand for better information retrieval has reignited the interest in Question Answering systems. Question Answering is considered as more complex NLP application than other types of applications like information retrieval (IR) or information extraction (IE), and it is some time regarded as supreme of IR/IE. Typically QA is supported by Natural Language Processing and IE [8]. The purpose of question answering (QA) is to detect and present to the user a valid answer to a question, rather than giving documents that may be topically related to the question or may contain the answer [9].

2.1 Basic Architecture of Question Answering System

The basic architecture of QA system consists of several components as shown (Figure 1) [10]. The role and responsibility of each component is explained below [11]:

1. **Query interface**: Query Interface is used to retrieve the question posted by the user.

2. **Query analyzer**: Query Analyzer phrase the question into subject, verb, object etc. It also used to improve the performance of the QA System.

3. **Question classification**: Question Classification is used to recognize the type of the question and after that the type of the answer will be specified.

4. **Query Reformulation**: Query Reformulation play vital role in QA system because this component is used to find the correct answer to user question.

5. **Search Engine Module:** Search Engine Module is used to get back the document based upon important keyword present in the question. This component also spotlights the pattern of the question to get back the relevant document.

6. **Answer Extraction:** Search engine send candidate answers collection to next answer extraction module which extract candidate answers from retrieved documents.

7. **Answering Filtering and Ordering:** The candidate answers pass to filtering and ordering unit. Based on co-occurrence of the words and semantic relations obtainable in database ontology, answer type and keywords which were extracted in question processing module, system filter candidate answers collection.

8. As a result some answers which are not related with the asked question will be eliminated. The answers with high priority are exposed to user for validation. Then the answers get a validation answer grade and save it in knowledge base. If the user accepts the suggested answer which system presented as an exact answer.

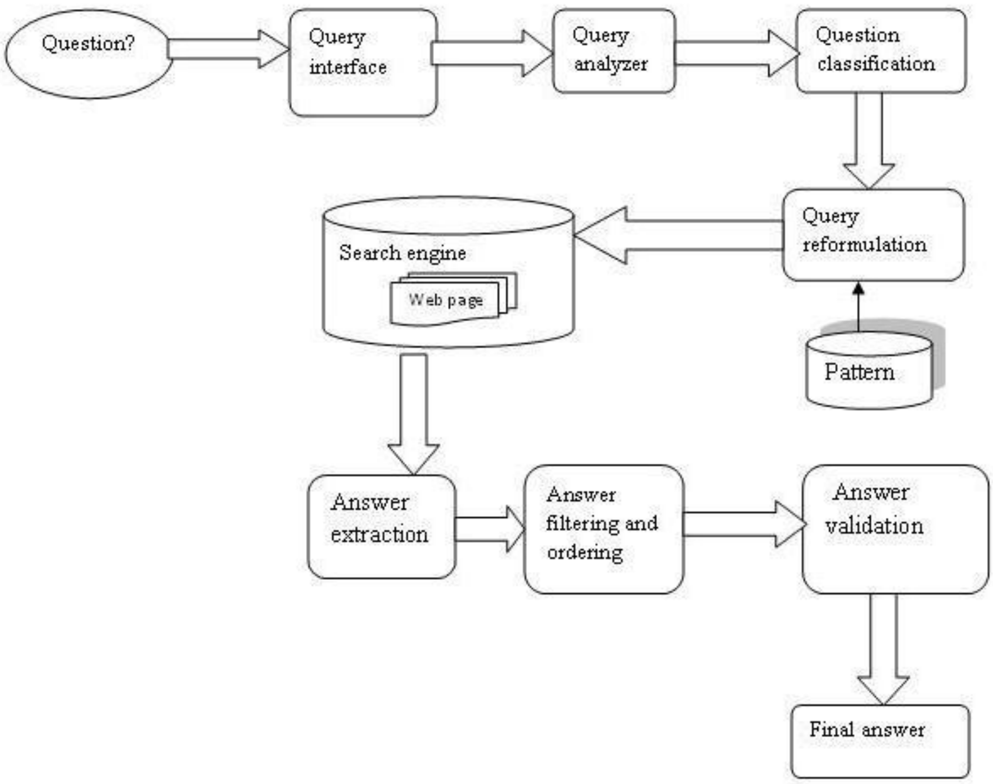

Fig 1: Question answering system architecture

2.2 Classification of Question Answering Systems

Question Answering Systems are classified on the basis of content and on the basis of language paradigm. Following are the different categories of Question Answering System [12]:

2.2.1 Classification based on Data Content

The most important classes of techniques based on data content are open-domain and restricted-domain. These two domains use word list and thesauri in classifying documents and classifying the questions.

2.2.1.1 Open Domain Question Answering (ODQA) System

Open domain question answering (ODQA) [13] deals with questions about nearly everything and can only depend on basic ontology. ODQA has become a very favorite research area over the past few years, due in large measure to the incentive of the TREC [14] Question Answering track. The TREC (Text Retrieval Conference) is a chain of workshops planned to progress the state of the art in text retrieval by giving the infrastructure necessary for large scale assessment of text retrieval methodologies. Ask Jeeves ("http://www.ask.com") is the well recognized ODQA system. Ask Jeeves is the fifth most popular search engine and the only natural language search engine on the list, according to the Nielsen/Net Ratings Mega

View Search report in 2005. Asking the question "Who is the president of the United States" to Ask Jeeves, extracts "The Chief of State of the United States is President George W. Bush, who is also Head of State" as the answer. START [15], AnswerBus [16], Brainboost [17], Yahoo Answering system [18], and Inferret [19] etc. are the different Question Answering Systems that are running on the web to accomplish user requirements.

2.2.1.2 Restricted Domain Question Answering (RDQA) System

Restricted-domain question answering (RDQA) [20] agrees with questions under a specific domain like medicine or railways. The likely questions are limited by the domain, hence it is possible to program entire domain knowledge in the system to evaluate questions or answer sources. In this RDQA, system gives accurate answer instead of giving a set of documents related to answer. The answer sources can be completely structured data, which is easy to process. Green's BASEBALL system is a one of the restricted-domain QA system that just answers questions about one season's baseball data.

2.2.2 Classification based on Language

We can represent QA systems by the source language, that represents questions, and target language, that represents answers. QA systems are classified as follows:

2.2.2.1 Monolingual QA System

Questions and answers are in the same language in Monolingual QA System. Monolingual QA useful for people speaking one of the popular languages, and researchers have emphasized a great deal of interest to monolingual QA research. Monolingual system can rely on as few resources as possible.

2.2.2.2 Cross lingual/Trans lingual QA System

The question is posed in a source language and the answer must be found in a target collection of a different language in Cross lingual QA (CLQA). In CLQA concentration is paid on translating the query into English and executing monolingual English QA on the translated query.

2.2.2.3 Multilingual QA System

User asks questions in one language and gets answers different from the source language or same as source language in Multilingual QA (MLQA) system. In designing such a system, there is a need to pay concentration on the linguistic side. Multilingual QA has come out only in the last few years as a corresponding research task, representing a perfect direction for at least two reasons. First, it permits users to communicate with machines in their own languages, contributing to simple, faster, and more information extraction. Second, cross lingual abilities enable QA systems to retrieve information stored only in language specific text collections.

3. JAVELIN III: CROSS-LINGUAL QUESTION ANSWERING FROM JAPANESE AND CHINESE DOCUMENTS

JAVELIN is a question-answering system with a prolongable architecture [21]. JAVELIN architecture is language-free architecture. In this architecture the English version of JAVELIN is extended for cross-language question answering between English and Chinese or Japanese. JAVELIN contributed in four CLQA subtasks (J-J, E-J, C-C, E-C), for which a total of 11 official and 6 unofficial runs were submitted. Best run for E-J achieved about 13% in answer accuracy and the best E-C run achieved 19% in answer accuracy.

3.1 Javelin III Architecture

The JAVELIN system consists of five main modules: Question Analyzer (QA), Translation Module (TM), Retrieval Strategist (RS), Information Extractor (IEX) and Answer Generator (AG). The QA module is parses the input question, choosing the valid answer type, and then giving a set of keywords. Keywords are translated to task-specific languages by Translation Module (TM). Answers to the question using translated keywords are contained in RS Module. Answers from the relevant documents are extracted by IEX module. Answers are normalized and ranked in order of their correctness by the AG module. The complete architecture is shown in Figure 2 [22].

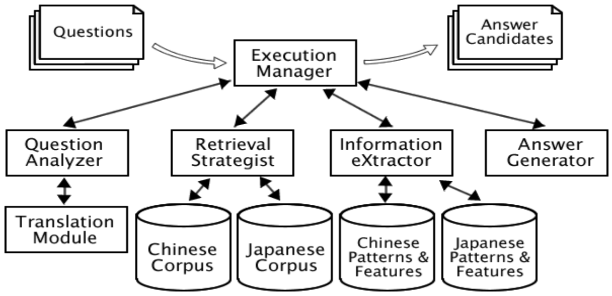

Fig 2: JAVELIN III Architecture

4. DIALOGUE BASED QUESTION ANSWERING SYSTEM IN TELUGU

For Railway information in Telugu a dialogue based Question Answering (QA) system has been illustrated. Telugu is a significant language in India. The major component of this QA system is the Dialogue Manager (DM) which is used to handle the dialogues between user and system. It is essential in producing dialogue for explaining partially understood questions, deciding Anaphora and Co-reference drawbacks. Different modules have been produced for handling the query and its conversion into formal database query language statement(s). Depending on the output from the database, a natural language answer is produced. The pragmatic results acquired on the current system are cheering. Trying with a set of questions in Railway domain, the QA system demonstrated 96.34% of precision and 83.96% of dialogue success rate.

4.1 Architecture of Dialogue based QA System in Telugu

In this keyword based approach the input query statement is analyzed by the query analyzer, which uses domain ontology stored as knowledge base, generating tokens and keywords. The appropriate query frame is selected based on the keywords and the tokens in the query statement. Each query frame is associated with a SQL generation procedure. The appropriate SQL statement(s) is generated using the tokens retrieved from the input query. The QA system architecture is shown in Figure 3 [23]. The Dialogue Manager keeps track of the elliptical queries from the user that constitute the dialogue and helps in the SQL generation procedure using dialogue history [24], which contains information about previous tokens and their types as well as other dialogue information like answers retrieved by the current SQL statements and the answers for previous queries in the dialogue. The SQL statements used to retrieve the correct answer from the database. Based on the result of the DBMS, a natural language answer is generated.

This answer is forwarded to the DM for onward transmission to the user.

The system enters into a dialogue with the user through the DM if the system is not able to decide on the query frame by using the keywords generated from the input query. An interactive message is delivered to the user through the DM during SQL generation if it is found that more information is required from the user to output the SQL statement. Then the user sends the required information to the system. DM sends an error message to the user if user could not give correct information. DM will send a combined message depending on the user query when the SQL statement gives a null response from the database.

4.1.1 Query Analyzer

In this phase, Morphological analysis of the input query statement is done to spot the root terms. The system recognizes several tokens such as Station name, Reservation class, date and period of the day etc. and a set of keywords. For recognizing these tokens and keywords the query analyzer checks with the knowledge base. There is a probability that some terms may not be present in the knowledge base. Those words are simply discarded.

Here query is broken based on spaces. After parsing of each word, it is examined in the knowledge base until the word is discovered. After searching each term in the knowledge base, their semantic information and types are arranged in a list of tokens. Each token has the token value, its type and semantic information. These keywords and tokens are used to decide the query frame.

4.1.2 Query Frame Decision

The keywords in the input query are determined during the analysis of query. Based on the tokens and keywords, the appropriate query frame is determined. The extent of the user request can be focused by limiting the query domain and information resource. There are a finite number of anticipated question topics. Each anticipated question topic is explained under a single query frame.

4.1.3 Dialogue Manager

DM controls the flow of dialogue by checking how the system should reply to a user request and the synchronization of the other components in the system. Clarification questions are identified by the DM and given to the user if some information is not present or a request is ambiguous. When the user is not able to give correct information then at that time DM produces an error message explaining that missed information. If user asks questions without knowledge then DM generates a cooperative message, which will help the user in future requests. DM makes use of the dialogue history if the information is not present. Dialogue history maintains the information regarding what has been talked in the past and what is talking at present. It is used for dialogue control of context dependent requests. The DM obtains a semantic frame from the other system components. The DM then checks for missing information or sends a SQL query. DM checks if new information is contained in the query or the information is conflicting to information given before the query is sent off. In that case DM can either keep the previous information or replace it with the new one. The accuracy of the system primarily depends on the characterization of the dialogue history and how the DM reacts to the user's dialogue.

4.1.4 SQL Generation

The corresponding procedure for the SQL query generation is called when query domain and information resource are restricted. There is a procedure for SQL statement(s) generation for each query frame. It needs the tokens produced by the query analyzer in order to generate the SQL query.

4.1.5 Answer Generation

When the SQL statement for an input query statement is produced, it is activated on the database and the produced information is used to denote the answer. The retrieved information is revised in the dialogue history for future reference.

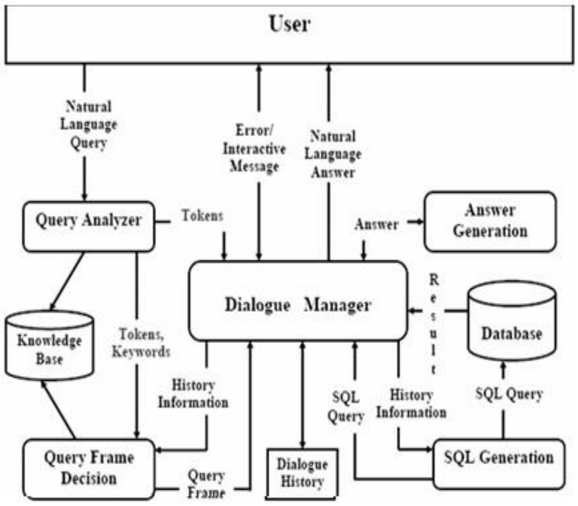

Fig 3: Dialogue Based QA System Architecture

5. JAPANESE QUESTION-ANSWERING SYSTEM USING A* SEARCH AND ITS IMPROVEMENT

In this system [30], several measures of the degree of sentential matching and an alternative of a voting method are introduced in order to improve the accuracy. Both can be incorporated into this system of controlled search. In this effectiveness of the newly introduced techniques at Subtask 1 of NTCIR4 QAC2 are examined, which is an evaluation workshop for question-answering systems in Japan.

5.1 Methods and its Limitations

Table 1. Methods of Japanese Question Answering System

Method	Purpose	Limitation	MRR(Mean Reciprocal Rank)
A* search control in a sentential matching	Reduce turnaround time while maintaining accuracy	Accuracy not sufficiently high.	0.3

Several measures of degree of sentence matching and a variant of voting method were integrated with A* search control method	To improve accuracy of A* search control method	Higher accuracy as compared to previous approach	0.5

5.2 Details of Experimental System

Experiments were conducted under the conditions as shown in Table 2. Each paragraph in the experiments consists of three sentences.

Table 2. Experimental System Details

Morphological Analyzer	JUMAN 3.61 [25]
Dependency Analyzer	KNP 2.0b6 [26]
NE recognizer	SVM-based NE recognizer [27]]using SVMlight [28]
Numerical expression extractor	System by Fujihata et al. [29]
Document database (knowledge resource) Size: (774 MByte)	Mainichi Shimbun newspaper articles in 1998 and 1999 Yomiuri Shimbun newspaper articles in 1998 and 1999
Computer	CPU: Xeon (2.2 GHz) × 2, Main memory:4 GByte (for QA server) CPU: UltraSPARC III Cu (900 MHz) × 2, Main memory:8 G Byte (for search engine)
Language of Implementation	JPerl 5.005_03

5.3 System Overview

An overview of the proposed system is shown in Figure 4. [30] It has the following features.

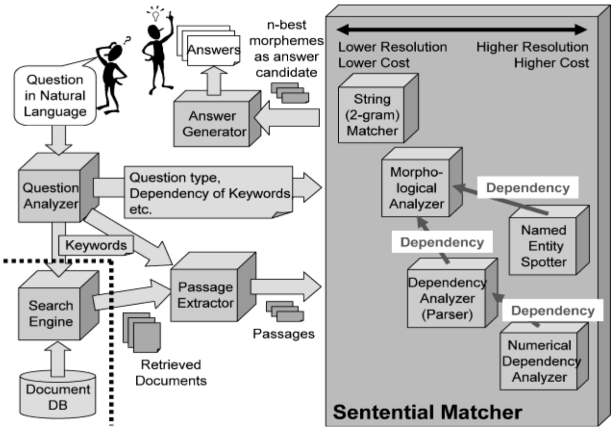

Fig 4: System Overview

5.3.1 Question Analyzer

The question analyzer accepts a question from a user and obtains the following information: (1) results of the morphological analysis and syntactic parsing, (2) a record of keywords, (3) the question type, and (4) dependency structures of numerical expressions.

5.3.2 Passage Extractor

Since the information associated to a question is mainly contained in a very small part of the document, the passage extractor divides each document into small passages and chooses suitable passages that are associated to keywords.

5.3.3 Sentential Matcher

The input for this module is a collection of sentences in retrieved passages or retrieved sentences. The module considers each morpheme as an answer candidate and allots it a matching score. The matching score denotes the fitness of answer candidate for the answer. Related to other QA systems, the score is computed using the following steps: (1) Connect the answer candidate AC to an interrogative Q in a question after assuming that it is an answer. (2) Following the previous condition, compute the matching score based on the similarity between the contexts of AC and Q .This module outputs a list of n-best morphemes with scores.

5.3.4 Answer Generation

An answer candidate achieved by the sentential matcher is a morpheme that may be a word or a part of a longer compound word. If it is a longer compound word then the system finds the compound word including the answer candidate and outputs it.

6. EVALUATION METRICS

There are several parameters that are used to analyze the performance of different Question Answering Systems. In this section we describe some of the evaluation metrics used in Question Answering System to evaluate its performance: [31]

6.1 R-Accuracy and RU-Accuracy

R-Accuracy and RU-Accuracy are used to evaluate Question Answering performance. A Question Answering system gives a list of ranked answer responses for every question, but R-accuracy and RU-accuracy only believes the correctness of the top 1 rank answer response on the list. An answer response is a pair that consists of an answer and its basis document. Each answer response is judged as Right, Unsupported, or wrong, as defined in the NTCIR-6 CLQA overview [32]:

Right (R): the answer is correct and the basis document holds it.

Unsupported (U): the answer is correct, but the basis document cannot support it as a correct answer. There is insufficient information in the document for users to confirm by themselves that the answer is the correct one.

Wrong (W): the answer is incorrect.

Based on this criterion, the accuracy is calculated as the no. of correctly answered questions divided by the total no. of questions. R-accuracy means that only right judgments are regarded as correct, while RU-accuracy means that both right and unsupported judgments are counted.

$$R\text{-}Accuracy = \frac{Total\ no.of\ questions\ for\ which\ the\ top1\ rank\ answer\ is\ Right}{Total\ no.of\ questions}$$

$$RU\text{-}Accuracy = \frac{total\ no.of\ questions\ for\ which\ the\ top1\ rank\ answer\ is\ Right\ or\ Unsupported}{Total\ no.of\ questions}$$

6.2 Mean Reciprocal Rank

MRR is used to measure the Question Answering performance based on all the top ranked answers, not presently top1 answer. The MRR is computed as follows [31]:

$$\text{MRR} = \frac{1}{Total\ no.of\ questions}$$

$$\sum_{question_i} \begin{cases} \frac{1}{the\ highest\ rank\ of\ correct\ answers}, \\ (if\ a\ correct\ answer\ exists) \\ 0. \\ (if\ no\ correct\ answer) \end{cases}$$

6.3 Expected Answer Accuracy

In addition to normal answering accuracy metrics, a new metric known as Expected Answer Accuracy is proposed. Some case is there where one method is not better than the other, but has higher accuracy or MRR value. This experience usually occurs when several top answers have the equal ranking score. Expected Answer Accuracy is used to solve such problems. EAA score of a ranking method is defined as follows [31]:

$$\text{EAA} = \frac{1}{Total\ no.of\ questions}$$

$$\sum_{question_i} \frac{total\ no.\ of\ correct\ answers\ with\ top1\ rank\ score}{total\ no.\ of\ answers\ with\ top1\ rank\ score}$$

6.4 Stability method

The soul of the stability method is to compare systems x and y in terms of metric M using B different topic sets and count how often x outperforms y, how often y outperforms x, and how often the two are regarded as equivalent [33]. Let B = 1000. Let x and y denote a pair of systems from S, and S denote a set of systems (i.e., runs) submitted to a particular task. Let Q denote the entire set of questions used in the task, and let c be a constant. Let $M(x,Q_i)$ denote the value of metric M for System x averaged over a topic set $Q_i(\subset Q)$. Then, using the algorithm shown in Figure 4, the minority rate (MR) and the proportion of ties (PT) of M, given a fuzziness value f, can be computed as [34]:

$$MR = \frac{\sum_{x,y\in S} \min\left(GT_M(x,y), GT_M(y,x)\right)}{B\sum_{x,y\in S}} \quad (1)$$

$$PT = \frac{\sum_{x,y\in S} EQ_M(x,y)}{B\sum_{x,y\in S}} \quad (2)$$

```
for each pair of runs x, y ∈ S
    for each trial i = 1 to B
        select Q_i ⊂ Q at random s.t. |Q_i| == c;
        margin = f * max(M(x, Q_i), M(y, Q_i));
        if( |M(x, Q_i) − M(y, Q_i)| < margin )
            EQ_M(x, y) + +
        else if( M(x, Q_i) > M(y, Q_i) )
            GT_M(x, y) + +
        else
            GT_M(y, x) + +;
```

Fig 4: Algorithm for computing EQ_M(x, y), GT_M(y, x).

From the algorithm, it is clear that $GT_M(x, y) + GT_M(y, x) + EQ_M(x, y) = B$ for each run pair. The minority rate is an approximate of the chance of reaching a wrong conclusion about a pair of runs using a given metric, while the proportion of ties reflects its lack of discriminative power. Thus, for a good performance metric, both of these values should be small.

6.5 Swap method

The soul of the swap method is to approximate the swap rate, which denotes the likelihood of the event that two experiments, using two entirely different sets of topics, are ambiguous given an overall performance difference in terms of metric M. This method can be used to contrast the sensitivity of different metrics. Let d designate a performance difference between two systems. The swap method starts by defining 21 performance difference bins, where the initial bin represents performance differences such that $0 \leq d < 0.01$, the second bin represents those that $0.01 \leq d < 0.02$, and so on, and the last bin represents $0.20 \leq d$. Let BIN (d) denotes a mapping from a difference d to one of the 21 bins where it go. The algorithm shown in Figure 2 computes the swap rate for each bin. Both the stability and the swap method use sampling with no replacement from the original topic set Q. No duplicate topics are allowed within each Q_i (and $Q\__i$). In addition, the swap method ensures that Q_i and $Q\__i$ are disjoint [33].

```
for each pair of runs x, y ∈ S
    for each trial i = 1 to B
        select Q_i ⊂ Q and Q'_i ⊂ Q such that Q_i ∩ Q'_i == φ and |Q_i| == |Q'_i| == c;
        d_M(Q_i) = M(x, Q_i) − M(y, Q_i);
        d_M(Q'_i) = M(x, Q'_i) − M(y, Q'_i);
        count(BIN(d_M(Q_i))) + +;
        if( ( d_M(Q_i) * d_M(Q'_i) > 0 ) then
            continue
        else
            swap_count(BIN(d_M(Q_i))) + +;
for each bin b
    swap_rate(b) = swap_count(b)/count(b);
```

Fig 2: Algorithm for computing the swap rates

Provided a "confidence level" (e.g., 95%), we can plan the swap rate (i.e., 1 minus the "confidence level") against the performance difference bins, so that the minimum difference necessary to guarantee that a system is better than another at that confidence level can be achieved .

7. ACKNOWLEDGEMENT

I would like to express my thanks to Mr. Vishal Gupta, Assistant Professor in UIET, Department of Panjab University Chandigarh for his guidance in accomplishing this task.

8. REFERENCES

[1] Gunawardena, T., Lokuhetti, M., Pathirana, N., Ragel, R., and Deegalla, S. 2010. An Automatic Answering System with template matching for Natural Language Questions. In Proceedings of IEEE ICIAfS.

[2] Ahn, C. M., Lee, J. H., Choi, B., and Park, S. 2010. Question Answering System with Recommendation using Fuzzy Relational Product Operator. In Proceedings of ACM iiWAS Conference on Intelligent Systems, France.

[3] E. Voorhees. 1999. The TREC-8 question answering track evaluation. In The Eighth Text Retrieval Conference (TREC-8), pages 83–106. Department of Commerce, National Institute of Standards and Technology.

[4] Kamps, J., Geva, S., and Trotman, A. 2008. Report on the SIGIR workshop on focused retrieval. SIGIR Forum, 42(2):59 – 65, 2008.

[5] Smucker, M. D., Allan, J., and Dachev, B. 2012. Human Question Answering Performance Using an Interactive Document Retrieval System. In Poceedings of ACM lliX. Netherlands.

[6] Woods, W. A., Kaplan, R. M., and Webber, B. N. 1972. The Lunar Sciences Natural Language Information System: Final Report. BBN Report 2378, Bolt Beranek and Newman Inc., Cambridge, Massachusetts.

[7] Green, W., Chomsky, C., and Laugherty K. 1961.BASEBALL: An automatic question answerer. In Proceedings of the Western Joint Computer Conference, 219-224.

[8] Fleischman, M., Hovy, E., and Echihabi, A. 2003. Offline strategies for online question answering: Answering questions before they are asked. ACL.

[9] Pakray, P. 2007. Multilingual Restricted Domain QA System with Dialogue Management ME Thesis. Jadavpur University of Kolkata.

[10] Ramprasath, M. and Hariharan, S. 2012. A Survey on Question Answering System. International Journal of Research and Reviews in Information Sciences. Vol. 2, No. 1.

[11] Brill, E., Dumais, S., and Banko, M. Ask MSR question answering system. Microsoft Research One Microsoft Way Redmond, Wa. 8052

[12] Kangavari, M. R., Golpour, M., and Ghandchi, S. 2008. World Academy of Science, Engineering and Technology.

[13] Yang, H., Chua, T. S., Wang, S., and Koh, C. K. 2003. Structured use of external knowledge for event based open domain question answering. *SIGIR*.

[14] Voorhees, E. M. 2004. Overview of the TREC 2003 question answering track. In Proceedings of the 12th Text Retrieval Conference.

[15] START Question Answering System. Website: http://start.csail.mit.edu.

[16] AnswerBus, Question Answering System. Website: http://answerbus.com

[17] BrainBoost, Question Answering Sytem. Website: http://www.answers.com/bb/

[18] Yahoo! Question Answering System. Website: http://answers.yahoo.com/

[19] [19] Inferret, Question Answring System. Website: http:/asked.jp

[20] Diekema, A. R., Yilmazel, O., and Liddy, E. D. 2004. Evaluation of Restricted Domain Question-Answering Systems. In Proceedings of the ACL2004 Workshop on Question Answering in Restricted Domain, 2-7.

[21] Nyberg, E., Frederking, R., Mitamura, T., Bilotti, M., Hiyakumoto, H., and Ko J., Lin, F., Lita, L., Pedro, V., and Schlaikjer, A. 2005. JAVELIN I and II systems at TREC 2005. In Proceedings of TREC.

[22] Mitamura, T., Lin, F., Shima, H., Wang, M., Ko, J., Betteridge, J., Bilotti, M., Schlaikjer, A., and Nyberg, E. 2008. School of Computer Science. Carnegie Mellon University

[23] Reddy, R. R. N. and Bandyopadhyay, S. 2006. EACL Workshop on Multilingual Question Answering. MLQA).

[24] Eriksson, F., Annika, and Arne, J. 2000. Dialogue and Domain Knowledge Management in Dialogue Systems. In Proceedings of 1st SIGDIAL workshop at ACL2000.

[25] Kurohashi, T. and Nagao, M. 1998. Japanese Morphological Analysis System JUMAN version 3.6 Manual. Kyoto University. Japan.

[26] Kurohashi, S. 1998. Japanese Syntactic Parsing System Knp Version 2.0 B6 Instruction Manual.

[27] Yamada, H., Kudo, T., and Matsumoto, Y. 2002. Japanese Named Entity Extraction Using Support Vector Machine. IPSJ Journal 43, 1 (Jan.), 44–53. Japan.

[28] Joachims, T. 2002. SVM*light*—Support vector machine. http://svmlight.joachims.org/.

[29] Fujihata, K., Shiga, M., and Andmori, T. 2001. Extraction of Numerical Expressions by Constraints and Default Rules of Dependency

Structure. Sig Notes 2001-Nl-145. Information Processing Society of Japan.

[30] Mori, T. 2005. Japanese Question Answering System Using A^* Search and Its Improvement. In ACM Transactions on Asian Language

Information Processing, Vol. 4, No. 3, Pages 280–304.

[31] Day, M. Y., Sung, L. C., Lee, Y. H., Jiang, T. J., Wu, C. W., Shih, C. W., Chen, Y. R., and Hsu, W. L. 2008. Boosting Chinese Question Answering with Two Lightweight Methods: ABSPs and SCO-QAT. In ACM Transactions on Asian Language Information Processing, Vol. 7, No. 4, Article 12.

[32] Lee, C. W., Day, M. Y., Sung, C. L., Lee, Y. H., Jiang, T. J., Wu, C. W., Shih, C. W., Chen, Y. R., and Hsu, W. L. 2007. Chinese-Chinese and English-Chinese question answering with ASQA at NTCIR-6 CLQA. In Proceedings of NII-NACSIS Test Collection for Information Retrieval Systems (NTCIR'07), Tokyo, Japan, 175–181.

[33] Sakai, T. 2007. On the Reliability of Factoid Question answering Evaluation. In ACM Transactions on Asian Language Information Processing, Vol. 6, No. 1, Article 3.

[34] Buckley, C. and Voorhees, E. M. 2000. Evaluating evaluation measure stability. In Proceedings of ACM SIGIR. 33–40. Language Information Processing, Vol. 7, No. 4, Article 12.