International Journal of Research in Computer Science
eISSN 2249-8265 Volume 5 Issue 1 (2015) pp. 15-26
www.ijorcs.org, A Unit of White Globe Publications

ADVANCEMENTS IN QUESTION ANSWERING SYSTEMS TOWARDS INDIC LANGUAGES

[1]Gursharan Singh Dhanjal, [2]Prof. Sukhwinder Sharma

[1]*M.Tech (IT), Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, INDIA*
Email: gursharan.info@gmail.com

[2]*Assistant Professor, CSE/IT Department, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib Punjab, INDIA*
Email: sukhwinder.sharma@bbsbec.ac.in

Abstract: *Limited work has been reported for automated Question and Answering (Q&A) systems in vernacular languages like Punjabi. This was validated from this systematic study on Q&A systems. This work discusses the typical architecture and issues/challenges associated in building multiple types of Q&A systems. Future directions as well as a new metric for selection of most probable answers have been discussed based on the limitations of previous Q&A systems.*

Keywords: natural language processing, question answering system, open domain qa, information retrieval

I. INTRODUCTION

Question Answering (Q&A) is a research field that is composed of various fields of Computer Science, Natural Language Processing (NLP), Information Retrieval (IR) and Information Extraction (IE). Currently, most Question Answer systems (search engines) provide relevant context on the basis of input questions. Because, what a user really wants is often a precise answer to a question. For instance, given the question "Who was the first American in space?", what a user really wants is the answer "Alan Shepard", but not to read through lots of documents that contain the words "first", "American" and "space" etc [1].

Research shows that people post questions to navigate the plethora of information through search engines. Go-gulf.com statistics suggests that 92% of user activities on the internet are search queries on search engine. People now usually search in the form of questions finding answers but search engines return only ranked lists of documents and don't deliver answers to the user. This problem is addressed by a question answering system. The best systems are now able to answer more than two third of factual questions [2]. Real world users of Q&A systems find them useful if they provide means to reach these goals. [3]:

- Timeliness
- Accuracy
- Usability
- Completeness
- Relevance

In the next section, this paper discusses the general architecture of Question Answering System. Third section contains classification of Q&A systems. Fourth section classifies the questioners and questions. Fifth section describes the challenges / issues in building a Q&A system. Sixth section describes the previous work(s) for Q&A systems. Seventh section discusses the research gaps found in this study and eighth section concludes this paper.

II. BASIC ARCHITECTURE

As shown in (Figure 1), a typical Q&A system consists of three modules. All of these modules have a core component beside other supplementary components. *Query Processing Module, Document Processing Module* and the *Answer Processing Module,* have Question Classification, Information Retrieval, and Answer Extraction as their core components respectively.

Question processing module identifies the prime focus of the question, and classifies the type of question. It also finds the answer type expected, and then devises multiple semantically equivalent tokens from the question.

Reformulation of a question (query expansion) into similar meaning questions boost up the recall of the information retrieval (IR) system, which is very important for question answering, because if no correct answers are present in a document, no further processing can be carried out to find an answer. Precision and ranking of candidate passages in an IR system can also affect the performance of question answering [4].

The final component in a Q&A system is the Answer Processing module. It is the distinguishing

feature between Q&A systems and the usual sense of text retrieval systems because its technology is an influential and decisive factor on question answering system for the final results. So, this process is deemed to be an independent module in the question answering systems [5].

Typically, the following scenario(s) occur in a Q&A system (also described graphically in Figure 1) [5]:

1. At first, the user posts a question to the Q&A system input.
2. Next, in the Question Processing Module, the inner component, *Question Analysis* determines the prime focus of the input question, which enhances the accuracy of Q&A system.
3. *Question Classification* plays a vital role in the Q&A system by identifying the question type and the expected answer type.
4. In *Question Reformulation*, the question is rephrased by expanding the query and passing it to the Document Processing Module.
5. The *Information Retrieval* component is used to retrieve the relevant documents based on the important keywords that appear in the question.
6. *Paragraph Filtering* component retrieves the relevant documents, filters and shorten them into small paragraphs expected to contain the answer.
7. Next, *Paragraph Ordering* is performed on these filtered paragraphs / tokens and passed to the Answer processing module.
8. On the basis of answer type and other recognition techniques, *Answer Identification* is performed.
9. To perform *Answer Extraction and Validation*, a set of heuristics may be defined so that only the relevant word or phrase (answer) is extracted.

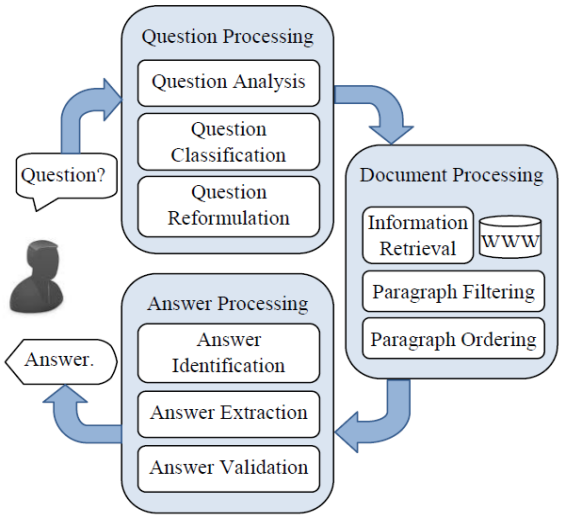

Figure 1: Typical Q&A System Architecture [5]

III. Q&A SYSTEMS CLASSIFICATION

A. *Domain Based Classification*

According to the application domain, Q&A systems are classified into two main categories:

1. Open domain Q&A Systems

In this type of Q&A systems, questions may be related to any subject, i.e. its domain is not restricted. The corpus may consist of unstructured or structured texts [6]. On the other hand, these systems tackle huge amount of data to extract the most relevant answer.

2. Restricted-domain Q&A Systems (domain specific)

Restricted domain means that the questions are restricted to a specific domain (music, sports etc. These types of Q&A Systems are easier to build, because the vocabulary is more predictable, and ontologies describing the domain are easier to construct. Only limited types of questions are accepted, such as questions asking for descriptive rather than procedural information [6].

B. *Approach Based Classification*

1. Linguistic Approach

This approach may contain AI based method using NLP technique and a corpus / knowledge base. Linguistic techniques such as tokenization, POS tagging and parsing may be implemented on user's question to formulate it into a precise query that merely extracts the respective response from a structured database. Building a knowledge base is a time-consuming process, so these systems are generally applied where problem is related to long-term information needs for a particular domain. The key limitation of these systems is that the knowledge stored in the structured database was only capable of answering questions asked within the restricted domain [7].

2. Statistical Approach

Importance of this approach is increased by rapid growth of online text repositories available. It is used where large amount of heterogeneous data is to be dealt with. Statistical approaches are independent of SQL and can formulate queries in natural language form. These approaches basically require an adequate amount of data for precise statistical learning but once properly learned, produce better results than other competing approaches [7]. An advantage of QA Systems built with Statistical Approach is that these are based on proven methods to find ranges of threshold, co-relation, and validity of the arguments. But there's also a disadvantage, as these systems may have a large sampling error. It is also said that co-relation does not guarantee the predictability of results. One of the pioneer works based on the statistical model was IBM's statistical QA System [8].

3. Pattern Matching Approach

Using the expressive nature of text patterns, this approach replaces the complex processing involved in other approaches. There are two approaches for Pattern Matching: *Surface Pattern based* and *Template based*. Most of the pattern matching Q&A systems, rely on surface text patterns while some of them also use templates for response generation. These days, many systems use learning of text patterns instead of using complex logics [7].

Table 1: Overall comparison between three approaches [7]

	Linguistic	**Statistical**	**Pattern Matching**
Question Type Handled	Factoid questions	Complex non-factoid along with factoids	Factoids, definition, acronym, birth date.
Heterogeneous data handling.	Quite difficult as knowledge base are generally designed only to handle their pre-stored data type.	Statistical similarity measurements are used to integrate data.	Easily possible as pattern aids in wrapper generation.
Semantic understanding	Deep	Shallow	Less than all other competing approaches.
Reliability	Most reliable as answers are extracted from self-maintained knowledge base.	Reliable as most of these systems use supervised approach.	Depends on the validity of knowledge resource.
Scalability	Quite complex as new rules have to be introduced in the knowledge base for every new concept.	Most suitable for handling large data once properly trained.	Less as new patterns have to be learned for each new concept.
Evaluation Technique/Test	Domain specific manually developed test collections.	TREC, CLEF, NTIRC test collections.	Domain specific manually developed test collections.
Application area	Systems that have long term information needs for specific domains	Quite suitable in handling large volume data e.g. web	Best suits to small and medium size websites, Semantic web.

IV. CLASSIFICATION OF Q&A SYSTEM ATTRIBUTES

A. Questioners

Complexity level of questions may vary according to the level of a questioner. The questioner may be a casual one, or a professional analyst, requiring a complex Q&A system to process the answers. But in the end all questioners have a common goal to get an accurate answer of the question. Different levels of Questioners are depicted in figure 2:

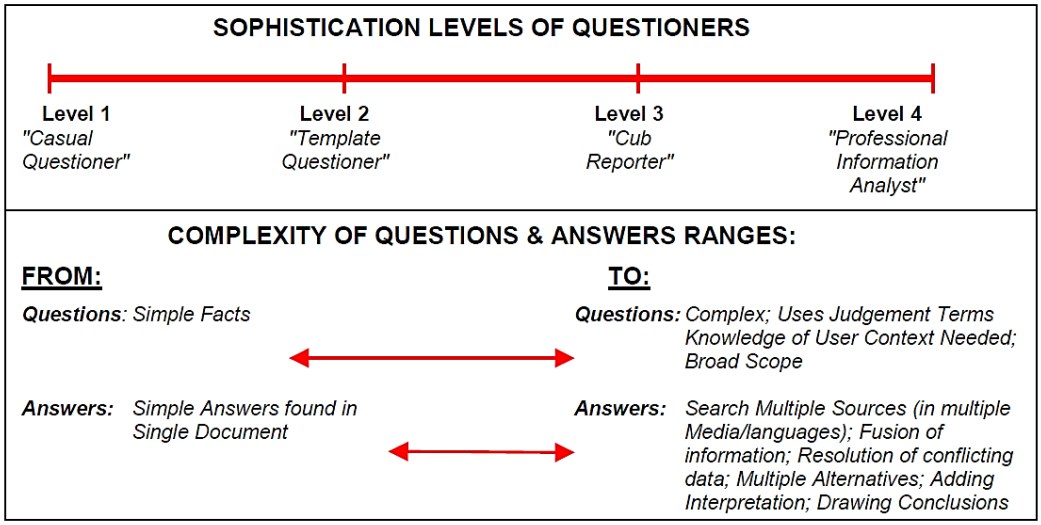

Figure 2: Complexity Levels of Questioners [3]

B. Question Classification

There are different types of methods available to classify the questions. QA research attempts to deal with a wide range of question types including functional, list, wh, non-wh, etc. Most of the relevant question types are discussed in following table.

Table 2: Comparison of different question classes

Question Category	Description	Pattern	Example
Functional Word Questions	All Non-Wh questions (except how) fall under this category.	These generally start with non-significant verb phrases	Name the first man to climb Mount Everest.
List Questions	A list question expects a list as its answer		Name the most-populated cities in world.
When Questions	When Questions starts with "When" keyword and are temporal in nature.	When (do|does|did|AUX) NP VP X	When did you complete your research work?
Where Questions	"Where Questions" starts with Where keyword and represent natural entities such as mountains, geographical boundaries, or any types of location (natural, manmade, virtual).	Where (do|does|did| AUX) NP VP X?	Where is Mount Everest located?
Which Questions	The expected answer type of such questions is decided by the entity type of the NP	Which NP X"?	Which Indian company is in the top 20 list?
Who/Whose/Whom Questions:	Here [word] indicates the optional presence of the term word in the pattern. These questions usually ask about an individual or an organization	(Who|Whose|Whom) [do|does|did|AUX] [VP] [NP] X?	Who was that guy? Whose purse is this?
Why Questions	Why Questions, always ask for certain reasons or explanations.	Why [do|does|did|AUX] NP [VP] [NP]" X"	Why is he upset?
How Question	For the first pattern, the answer type is the explanation of some process while second pattern return some number as a result	''How Questions" have two types patterns ''How [do/does/did/AUX] NP VP X?" or ''How [big|fast|long|many|much|far] X?"	How does he perform in the test? How are you?

V. CHALLENGES AND ISSUES IN BUILDING Q&A SYSTEM

In 2002, a group of researchers wrote a detailed roadmap of research in question answering, identifying the issues and challenges in building a Q&A systems [3].

1. *Question classes:* Different types of questions require different strategies to find an appropriate answer.
2. *Question Processing:* There are various ways (interrogative, assertive) to present a question with the same information request. This creates a problem of being understood as two different questions. A semantic model would recognize similar questions, regardless of how they are presented.
3. *Context and Q&A:* Questions are usually asked within a specific context and answers accordingly. To resolve ambiguities in question, context can be used by the Q&A systems.
4. *Data sources for Q&A:* It must be known beforehand, what knowledge sources are available and are relevant to the question. If the knowledge base / data sources, doesn't contain the answer to a question, no matter how well programmed the system is, a correct result is difficult to obtain.
5. *Answer Extraction:* Answer extraction depends upon the complexity of the question, on the answer type provided by question processing, on the actual data where the answer is searched, on the search method and also on the question focus and context.

6. *Answer formulation:* The result of a Q&A system should be presented in a way as natural as possible. For example, when the question classification indicates that the answer type is a name, a quantity or a date, the extraction of a single datum is sufficient. For other cases, presentation of the answer may require to combine the partial answers from multiple documents.
7. *Real time question answering:* There is a need for developing Q&A systems that are capable of extracting answers from large data sets in several seconds, regardless of the complexity of the question, the size and multitude of the data sources or the ambiguity of the question.
8. *Multilingual (or cross-lingual) question answering:* This is the ability to answer a question posed in one language using an answer corpus in another language (or even several). This allows users to consult information that they cannot use directly.
9. *Interactive Q&A:* It is often the case that the question processing part may fail to classify properly the question or the information needed for extracting and generating the answer is not easily retrieved. In such cases, the questioner might want not only to reformulate the question, but also to have a dialogue with the system.
10. *Advanced reasoning for Q&A:* More sophisticated questioners expect answers that are outside the scope of written texts or structured databases. To upgrade a Q&A system with such capabilities, it would be necessary to integrate reasoning components operating on a variety of knowledge bases, encoding world knowledge and common-sense reasoning mechanisms, as well as knowledge specific to a variety of domains.
11. *Information clustering for Q&A:* Information clustering for Q&A systems is a new trend that originated to increase the accuracy of Q&A systems through search space reduction.
12. *User profiling for Q&A:* The user profile captures data about the questioner, comprising context data, domain of interest, reasoning schemes frequently used by the questioner, common ground established within different dialogues between the system and the user, and so forth. The profile may be represented as a predefined template, where each represents a different profile feature.

VI. RELATED WORK

A. Previous Q&A Systems

Moldovan et al. (LASSO) [9], used NLP to find answers in large collections of documents. Question processing included the combination of syntactic information with semantic information, and further eight heuristic rules extracted the keywords used for identifying the answer. This research also introduced paragraph indexing, where retrieved documents were first filtered into paragraphs and then ordered.

Riloff and Thelen [10], developed a rule-based system, Quarc, that can read a short story and find the sentence in the story that best answers to the given question. Quarc uses the heuristic rules that look for lexical and semantic clues in the question and the story.

Tiansi Dong et al. [11], developed a Q&A system for the German language, aimed at providing concise and correct answers to arbitrary questions called *LogAnswer*. The paper presents a machine learning solution to the wrong answer avoidance (WAA) problem, applying a meta-classifier to the output of simple term-based classifiers and a rich set of other WAA features. It is designed as an embedded AI system which integrates methods from several fields of AI, namely NLP, machine learning, knowledge representation and automated theorem proving.

Gaizauskas and Humphreys (QA-LaSIE) [12], presented a system that performed linguistic analysis with an IR system linked with an NLP system. The IR formulated a question query to obtain set of ranked documents or passages. The NLP system presented semantic representation for each of the returned documents or passages.

Radev et al. (NSIR) [13], presented a probabilistic method for Web-based Natural Language Question Answering, called Probabilistic Phrase Re-ranking (PPR). Their NSIR web-based system utilized a flat taxonomy of 17 classes, in which two methods were used to classify the questions; the machine learning approach using a decision tree classifier, and a heuristic rule-based approach.

Li & Roth [14], contributed a hierarchical taxonomy in which questions were classified and answers were identified based upon that taxonomy. A machine learning technique called SNoW was tested in order to classify the questions into coarse and fine classes of the taxonomy. They also showed through another experiment, the differences between a hierarchical and flat classification of a question.

Ravichandran & Hovy [15], 2002, presented a method that learns patterns from online data using some seed questions and answer anchors, without needing human annotation. It exploited surface text information using manually constructed surface patterns for finding answers.

Zhang and Lee [16], worked on the limitation of the research by Li & Roth [14], and carried out a comparison between the following algorithms of machine learning: *Support Vector Machine (SVM), Nearest Neighbors (NN), Naïve Bayes (NB), Decision Tree (DT) and Sparse Network of Winnows (SNoW).* They have shown that SVM with a tree kernel can achieve performance improvement over a single-layer SNoW classifier using the same primitive syntactic features.

RJ Mammone et al. [8], describes the IBM Statistical Question Answering for TREC-9 system in detail and look at several examples and errors. The system is an application of maximum entropy classification for question/answer type prediction and named entity marking. The system retrieves document from local encyclopedia, expands query words and retrieves passage from TREC collection. An answer selection algorithm determining the best sentence is also presented.

Lita et al. [17], introduced, cluster-based query expansion method that learns queries known to be successful when applied to similar questions. Cluster-based expansion improves the retrieval performance of a statistical Q&A system when used in addition to existing query expansion methods. Paper shows that documents retrieved using the cluster-based approach are inherently different from the documents retrieved using existing methods and provide a higher data diversity to answers extractors.

Xu et al. [18], they adopted a hybrid approach that used various complementary components including information retrieval and various linguistic and extraction tools such as name finding, parsing, co-reference resolution, proposition extraction, relation extraction and extraction of structured patterns.

Peng et al. [19], presented an approach to handle the main limitations of work by Ravichandran & Hovy [15]. They explored a hybrid approach for Chinese definitional question answering by combining deep linguistic analysis (e.g. parsing, co-reference, named-entity) and surface pattern learning in order to capture long-distance dependencies in definitional questions.

Table 3: Comparison of previous Q&A Systems (derived from [5])

	Question Processing			**Document Processing**			**Answer Processing**				
Q&A Components / **Q&A Research**	Question Analysis	Question Classification	Question Reformulation	Information Retrieval	Paragraph Filtering	Paragraph Ordering	Answer Identification	Answer Extraction	Answer Validation	Performance	Limitation(s)
Moldovan et al (LASSO) **[9]**	✓	✓	✓	✓	✓	✓	✓	✓		MRR (strict)	Answer was correct only if it was among top 5 ranked long answers
										Short Answer: 55.5% / **Long Answer: 64.5%**	
Riloff and Thelen2000 **[10]**	✓	✓		✓			✓			Accuracy (40%)	Accuracy only 40%, given the simplicity of system rules.
Gaizauskas & Humphreys (QA-LaSIE) **[12]**		✓		✓			✓	✓		Short answers: Precision: 26.67% / Recall: 16.67%; Long answers: Precision: 53.33% / Recall: 33.33%	Success was limited, as only 2/3 of the test set questions were parsed
Radev et al. (NSIR) **[13]**		✓		✓	✓		✓	✓		MRR 20%	- Sentence segmentation, POS tagging and text chunking gave low results. - User submitted query is not reformulated
Li & Roth **[14]**		✓					✓			Accuracy (89%)	- No reason given to choose SNoW. - No other machine learning classifier tested which could've achieved better results

Ravichandran & Hovy **[15]**		✓		✓			✓	✓		Average MRR TREC Docs: 36.66% On Web: 56.66%	- Performed badly with general definitional questions, since the patterns did not handle long-distance dependencies - Only worked for certain types of questions that had fixed anchors, such as "where was X born"
Zhang and Lee **[16]**		✓		✓			✓	✓		Accuracy (SVM Algo.) Fine Grain: 80.2% Coarse Grain: 85.8%	Uses only automatically constructed features
Tiansi Dong et al. **[11]**	✓	✓	✓	✓			✓	✓		Wrong Answer Avoidance 90.6%	- With elimination of wrong answers, half of the correct answers are lost, with final answer rate of 3.7%. - Each answer candidate is to be examined manually against all other candidates, which is an overhead if there are large number of candidates
RJ Mammone et al. **[8]**		✓		✓			✓	✓		Error Rate (250 Byte): 43.8%	Doesn't handle semantic change properly.
Lucian Vlad Lita and Jaime Carbonell **[17]**	✓			✓			✓	✓		Precision 71%	Learning structure require more training questions for each cluster
Xu et al. **[18]**		✓		✓			✓	✓		F-Score 0.49 (Baseline)	Only "what" and "who" questions were tested.
Peng et al. **[19]**		✓		✓			✓	✓		Precision (learned patterns) who-is: 65.42% what-is: 68.58%	- Captured only limited syntactic information without providing any semantic information. - Simple POS Tagging used

B. Q&A Systems for Indic Languages

Kumar et al. [20], focused on developing Hindi Q&A system having different language constructs, query structure, common words. Primary goal was to help elementary and high school students in getting correct answers for their subject related questions. Self-constructed lexical database of synonyms was used, as a Hindi Word Net was not available. Case based rule classifies the question, change to proper query and submitted to retrieval engine.

Sahu et al. [21], discusses an implementation of a Hindi Q&A system "PRASHNOTTAR". The parser gives the analysis of a sentence in terms of morphological analysis, POS tagging and chunking. It presents four classes of questions namely: "when", "where", "what time" and "how many" and their static dataset includes 15 questions of each type.

Sekine et al. [22] developed a cross-lingual question-answering (CLQA) system for Hindi and English, which accepts questions in English, finds candidate answers in Hindi newspapers, and translates the answer candidates into English along with the context surrounding each answer. Initially, the examiner examined the questions and searched their answers from Hindi newspapers. An English Hindi bilingual dictionary was used to find out the top 20 Hindi articles, which were used to find out the candidate answers and finally, Hindi answers were returned back to the English language.

Banerjee et al. [23] developed a Bengali Question Classification System. Bengali is an important eastern Indic language. First step in developing a Q&A system is to classify natural language question properly. Their own work is extended to create a two-layer taxonomy is proposed with 9 course-grained classes & 69 fine-grained classes. The proposed automated classification work uses ensemble of multiple models. It is shown that boosting approach shows slightly better accuracy than bagging approach.

Reddy et al. [24], describes a dialogue based Q&A system in Telugu language for railway specific

domain. Main module of the system was dialogue manager, responsible for handling the dialogues between a user and system. This architecture was based on the keyword approach in which query analyzer generates tokens and keywords with the use of knowledge base. Based on the keywords and tokens, an appropriate frame was selected. The words that have some semantic information were needed to be present in the knowledge base. SQL statements were generated from the tokens. The railway database had been constructed which contained the information about the arrival / departure time of each train, information regarding their fares using a relational model. Basic responsibility of the dialogue manager was to manage the flow of dialogues. Dialogue manager was also responsible for the coordination of the other components in the system. After the generation of the query frame, SQL query was generated. Then the answer was retrieved from the database using SQL query.

Pakray [6], presented a system in which the query is specified through user, by starting a dialogue with the system. Shallow parser takes the input question, semantically tag it with the help of domain ontology. Tagged words were divided into chunks and with keywords present in the chunks, frame of the query was determined. Dialogue manager was used to obtain the missing information from the user query. SQL statements were generated corresponding to the query frame. With the help of SQL statements, answers were extracted from the database. The natural language answer was generated by the answer generator.

Stalin et al. (2012) [25] discussed the web based application for extraction of answers for a question posed in Hindi language from Hindi text. If the answer was not present in the Hindi text then the answers were searched on Google. This paper proposed a QA architecture that used words of sentence (question). The architecture of the system involved various modules, Query interface, Question classifier, Query formulation. It required the knowledge of the pattern of the question. Database sent all the candidate answers to the next module which was responsible for the extraction of the answers from the retrieved documents. Then all the candidate answers were displayed on the screen.

Gupta et al. [26] developed a Punjabi QA algorithm which uses a new hybrid approach to recognize most appropriate answers from multiple set of answers for a given question. The relevant answers are retrieved for different types of questions like: ਕਦੋਂ "when", ਕੀ "what", ਕੋਣ "who", ਕਿਉ "why" and ਕਿੱਥੇ "where". The identification or the classification of these questions is done with the help of stop word removers, stemmers [27] [28], regular expressions, scripts, and algorithm specific to the Question & Answer patterns [29]. This approach used pattern matching along with mathematical equations for both, identification of question as well as possible answers. A scoring system is built which checks the term frequency of the various possible nouns, verbs, ad verbs and possible answers for the given question.

Table 4: Comparison of Indic Language Q&A Systems

Research Work			No. of questions tested	Performance	Limitations
Praveen Kumar et al. [20]			150	Accuracy (59%)	Self-constructed lexical database limits the knowledge data, makes unsuitable for large systems
Shriya Sahu et al. [21]			60	Accuracy (68%)	Lack of semantic approach and dynamic data set
Rami Reddy et al. [24]			95	Precision (96.34%) Dialogue Success Rate (83.96)	Low dialogue success rate due to insufficient coverage of the domain
Shalini Stalin et al. [25]			Sets of 20	Inconclusive results are shown.	Inconclusive results are shown.
Somnath Banerjee et al. [23]			1100	Accuracy (87.63%)	Classification only limited to Bengali language
S. Sekine et al. [22]			56	MRR (25%)	-Machine Translation is prone to errors -Finds answers from Hindi Newspapers, but aimed at English speaking users
Partha Pakray [6]	Without Dialogue Management	Bengali	70	Precision (87.50%) Recall (80%)	-Low dialogue success rate due to insufficient coverage of the domain and considering fewer database

		Telegu	132	Precision (97.63%) Recall (93.93%)	tables -Sometimes the system is unable to obtain chunks correctly from the input query even if it had identified the right query frame. -Misinterpretation of dialogue history is also another problem.
	With Dialogue Management	Bengali	58	Precision (83.67%) Dialogue Success Rate (72.91%)	
		Telegu	62	Precision (96.49%) Dialogue Success Rate (89.06%)	
Poonam Gupta et al. [26]			200	Average Precision (85.66%) Average Recall (65.28%) MRR (0.43%)	-Domain oriented -Lesser coverage of question types

VII. DISCUSSION

Each approach used to build a Q&A system has its own pros and cons. Traditional Q&A systems could only tackle the current issues of question answering to a limited extent. But, usage of a combination of various approaches (hybrid approach) can be useful to the domain of Q&A. After a detailed study of different types of Q&A systems in English as well as of Indic Languages, It was found that only a few language resources are available for Indic languages and more so in case of Indic language Q&A systems. Only few number of research work(s) are present and the ongoing research work(s) in this context are not up to the current level of research progress and requirements. The prime reason of this is the complex nature of Indic Languages, in relation to computing. As per our study, we were able to find only a single Q&A system, described in notable research repositories, which was developed by Gupta et al. [26]. We have found the following research gaps in the previous systems that could be filled to develop a better QA system in Punjabi.

- Previous Q&A systems were not able to handle a vernacular language (like Punjabi) in a proper way.
- Previous systems are not able to work properly on unstructured documents as well as discrete data.
- Neither the systems are open (available for reuse/open source/reproducible), nor they have a scalable architecture of Q&A.
- Limited work based on distance based similarity measures have been reported in previous systems.
- Most of the systems did not work with both kinds of variables: a) variables that have co-relation, b) variables that do not have any co-relation.
- Previous systems are not taking advantage of the principles of sequence mining of the questions & answers and their pattern frequency.
- Many of them are not using principle of statistical significance test as well as on statistical distance measures at the same time.
- Many previous systems do not take care of calculations related to gunning fog measure, readability index, keyword density, fog index, concordance etc.
- A proper Punjabi Stemmer like porter stemmer algorithm is not used.
- Limited sized stop word dictionary is used.
- Punjabi POS taggers used for Q&A, were of limited capability.
- Most noted Punjabi Q&A systems did not include the most prominent question type, how (ਕਿਵੇਂ, ਕਿਸ ਤਰ੍ਹਾਂ) and its further combinations.
- Punjabi Semantic Object dictionaries were not used in the previous Q&A system.
- Previous Punjabi Q&A systems are not Web based, so it is not available to be used by everyone.
- Existing system used a scoring system which creates a frequency or inverse frequency tables. An answer with least distance from the question is given as most appropriate answer. This approach has been fairly successful in terms of accuracy, however there is an ample scope to enhance, extend this work. After this systematic study, it is recommended to develop a new metric for selection of most likely answers to a given question. The next section proposes a new metric that incorporates distance measures along with pattern discovery/matching methods using sequence mining.

A. Center of Gravity Metric (CoGM)

The CoG Metric is based on the concept of physics known as Center of Gravity. The general definition of Center of Gravity (CoG) is the source of power that provides moral or physical strength, freedom of action, or will to act, thus, the center of gravity is usually seen as the source of strength. The center of gravity (CoG) of an object is the average location of its weight. In physics, the center of gravity of an object is a point at

which the object's mass can be assumed, for many purposes, to be concentrated [30].

If (x_{cg}, y_{cg}) are the coordinates of the CoG of a collection of point masses m_1, m_2 , etc, located at coordinates $(x_1, y_1), (x_2, y_2)$ respectively, then:

Solving for the *x* -coordinate of the CoG:

$$Y_{cog} = \frac{\sum_{i=1}^{n} m_i y_i}{\sum_{i=1}^{n} m_i}$$

Similarly, the y-coordinate of the CoG is:

$$X_{cog} = \frac{\sum_{i=1}^{n} m_i x_i}{\sum_{i=1}^{n} m_i}$$

To understand the center of gravity method, let's take an example, imagine there are three neighborhoods: Applegate, Barstow, and Canterberry. We might draw a grid over a map of the area with horizontal and vertical coordinates, and find Applegate is centered at (5,3) meaning 5 coordinate points horizontally and 3 coordinate points vertically. Barstow is centered at (4,2) and Canterberry is centered at (3,6). The average of the horizontal, also called the x-axis, coordinates is (5+4+3)/3 = 4.0. So, we might assume that it would be good to centrally locate the fitness club we would put it at 4.0 on the horizontal axis. However, this fails to consider that there may be many more customers coming from one neighborhood than from another. Imagine that Applegate contains 200 potential customers, Barstow contains 75, and Canterberry contains 25. We would probably want to be nearer to Applegate than to Canterberry. Therefore, we use the customer counts as weights, recognizing that of 300 total potential customers, 200/300 = 66.6% come from Applegate. The weighting can be accomplished by multiplying each coordinate value by the corresponding customer forecast, summing across all customer locations, and dividing by the sum of the customer forecasts. For this example, the optimal horizontal coordinate is [31]:

$$X = \frac{(5 \times 200) + (4 \times 75) + (3 \times 25)}{(200 + 75 + 25)} = 4.58$$

Therefore, the optimal horizontal coordinate is 4.58 on the grid. So also, the vertical (y) coordinate can be calculated:

$$Y = \frac{(3 \times 200) + (2 \times 75) + (6 \times 25)}{(200 + 75 + 25)} = 3.00$$

So, the coordinates (4.58, 3.00) would give you the CoG of the location area. Similarly, this CoG metric can be applied on QA sets to calculate their center of gravity based on their record location. These values would further help in ranking the probable answers of the questions given.

VIII. CONCLUSION & FUTURE WORK

The interest in study and development of automated Q&A systems has grown rapidly in recent years. We believe that we are in an era which might bring divergent changes in Q&A domain, thus it can be considered as a golden era for Q&A. In this study paper we have provided a comparative view of Q&A systems for general languages as well as Indic Languages. This review paper also describes the different question answering approaches and different types of Q&A technologies like basic pattern matching, statistical analysis, artificial intelligence etc. We have also analyzed Q&A systems for Indic Languages and found a research gap to be filled by replacing the traditional pattern matching technique in the answer scoring system, with a new intelligent algorithm based on CoG Metric (CoGM). For future direction, it is suggested that we may develop a Punjabi QA System using the Structured Equation Modeling Method (SEM).

IX. REFERENCES

[1]. Dell Zhang and Wee Sun Lee, "A Web-based Question Answering System," Massachusetts Institute of Technology (DSpace@MIT), 2003.

[2]. L HIRSCHMAN and GAIZAUSKAS R., "Natural language question answering: the view from here," Natural Language Engineering, vol. 7, no. 4, pp. 275-300, 2001. doi: 10.1017/S1351324901002807

[3]. John Burger et al., "Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)," in Document Understanding Conferences Roadmapping Documents, 2001, pp. 1-35.

[4]. Svetlana Stoyanchev, Young Chol Song, and William Lahti, "Exact Phrases in Information Retrieval for Question Answering," in 2nd workshop on Information Retrieval for Question Answering (IR4QA), Manchester, UK, 2008, pp. 9-16.

[5]. Ali Mohamed Nabil Allam and Mohamed Hassan Haggag, "The Question Answering Systems: A Survey," International Journal of Research and Reviews in Information Sciences (IJRRIS), vol. 2, no. 3, September 2012.

[6]. Dr. Partha Pakray, "Multilingual Restricted Domain QA System with Dialogue Management," Master's Thesis Report, Jadavpur University, Kolkata, 2007.

[7]. Sanjay K Dwivedi and Vaishali Singh, "Research and reviews in question answering system," in International Conference on Computational

Intelligence: Modeling Techniques and Applications (CIMTA), 2013, pp. 417-424. doi: 10.1016/j.protcy.2013.12.378

[8]. Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, Adwait Ratnaparkhi, and Richard J Mammone, "IBM's Statistical Question Answering System," in Proceedings of the Text Retrieval Conference TREC-9, 2000.

[9]. D. Moldovan, "Lasso: A Tool for Surfing the Answer Net," in Proceedings of the Eighth Text Retrieval Conference (TREC-8), 1999.

[10]. Ellen Riloff and Michael Thelen, "A Rule-based Question Answering System for Reading Comprehension Tests," in ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, 2000. doi: 10.3115/1117595.1117598

[11]. Tiansi Dong, Ulrich Furbach, Ingo Glockner, and Bjorn Pelzer, "A natural language question answering system as a participant in human Q&A portals," in Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, 2011, pp. 2430-2435. doi: 10.5591/978-1-57735-516-8/IJCAI11-405

[12]. R. Gaizauskas and K. Humphreys, "A Combined IR/NLP Approach to Question Answering Against Large Text Collections," in Proceedings of the 6th Content-based Multimedia Information Access (RIAO-2000), 2000.

[13]. Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu, and Amardeep Grewal, "Probabilistic Question Answering on the Web," Journal Of The American Society For Information Science and Technology, vol. 56, no. 6, pp. 571-583, 2005. doi: 10.1145/511446.511500

[14]. Xin Li and Dan Roth, "Learning question classifiers," in Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), 2002, pp. 556-562. doi: 10.3115/1072228.1072378

[15]. Deepak Ravichandran and Eduard Hovy, "Learning Surface Text Patterns for a Question Answering System," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002, pp. 41-47. doi: dx.doi.org/10.3115/1073083.1073092

[16]. Dell Zhang and Wee Sun Lee, "Question Classification using Support Vector Machines," in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, 2003, pp. 26-32. doi: 10.1145/860435.860443

[17]. Vlad Lucian Lita and Jamie Carbonell, "Cluster-Based Query Expansion for Statistical Question Answering," in IJCNLP, 2008, pp. 426-433. doi: 10.1145/1571941.1572058

[18]. Jinxi Xu, Ana Licuanan, and Ralph Weischedel, "TREC2003 QA at BBN: Answering Definitional Questions," in TREC, 2003, pp. 98-106.

[19]. Fuchun Peng, Ralph Weischedel, Ana Licuanan, and Jinxi Xu, "Combining deep linguistics analysis and surface pattern learning: A hybrid approach to Chinese definitional question answering," in Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 307-314. doi: 10.3115/1220575.1220614

[20]. P. Kumar, S. Kashyap, A. Mittal, and S. Gupta, "A Hindi Question Answering system for E-learning documents," in Third International Conference on Intelligent Sensing and Information Processing, 2005, pp. 80-85. doi: 10.1109/ICISIP.2005.1619416

[21]. Shriya Sahu, Nandkishor Vasnik, and Devshri Roy, "Prashnottar: A Hindi Question Answering System," International Journal of Computer Science and Information Technology (IJCSIT), vol. 4, no. 2, pp. 149-158, 2012.

[22]. Satoshi Sekine and Ralph Grishman, "Hindi-English Cross-Lingual Question-Answering System," ACM Transactions on Asian Language Information Processing, vol. 2, no. 3, pp. 181-192, September 2003. doi: 10.1145/979872.979874

[23]. Somnath Banerjee and Sivaji Bandyopadhyay, "Ensemble Approach for Fine-Grained Question Classification in Bengali," in 27th Pacific Asia Conference on Language, Information, and Computation, 2013, pp. 75-84.

[24]. Rami Reddy Nandi Reddy and Sivaji Bandyopadhyay, "Dialogue based Question Answering System in Telugu," in EACL 2006 Workshop on Multilingual Question Answering - MLQA06, 2006, pp. 53-60. doi: 10.3115/1708097.1708108

[25]. Shalini Stalin, Rajeev Pandey, and Raju Barskar , "Web Based Application for Hindi Question Answering System," International Journal of Electronics and Computer Science Engineering, vol. 2, no. 1, pp. 72-78, 2012.

[26]. Poonam Gupta and Vishal Gupta, "Hybrid Approach for Punjabi Question Answering System," Advances in Intelligent Systems and Computing, vol. 264, pp. 133-149, 2014. doi: 10.1007/978-3-319-04960-1_12

[27]. Mandeep Singh Gill, Gurpreet Singh Lehal, and Shiv Sharma Joshi, "Part-of-Speech Tagging for Grammar Checking of Punjabi," The Linguistics Journal, vol. 4, no. 1, pp. 6-22, May 2009.

[28]. Vishal Gupta and Gurpreet Singh Lehal, "Automatic Keywords Extraction for Punjabi Language," International Journal of Computer Science Issues, vol. 8, no. 5, September 2011.

[29]. Vishal Gupta and Gurpreet Singh Lehal, "Automatic Text Summarization System for Punjabi Language," Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 3, August 2013. doi: 10.4304/jetwi.5.3.257-271

[30]. Hasan H. Khaleel, Rahmita O.K. Rahmat, DM. Zamrin, Ramlan Mahmod, and Norwati Mustapha, "Vessel Centerline Extraction Using New Center of Gravity Equations," International Journal of Computer Science, vol. 40, no. 2, February 2013.

[31]. Dr. Scott Sampson. (2012) Quantitative Tools for Service Operations Management. [Online]. HYPERLINK "http://services.byu.edu/t/quant/loc.html"

How to cite

Gursharan Singh Dhanjal, *Sukhwinder Sharma*, "Advancements in Question Answering Systems Towards Indic Languages". *International Journal of Research in Computer Science*, 5 (1): pp. 15-26. October 2015.