Recent Trends in Natural Language Question Answering Systems: A Survey

[1]Vaibhav Mishra, [2]Dr. Nitesh Khilwani
[1]Research Scholar, [2]Technical Architect
[1]Mewar University,
[2]Edifecs RoundGlass

Abstract - **The need to inquiry data content accessible in different organizations including organized and unstructured information has turned out to be progressively significant. Henceforth, Question Answering Systems (QAS) are essential to fulfill this kind of need. The QA systems are focused towards giving relevant answers in light of inquiries proposed in natural language. Question Answering system is an vital research area in IR. Research on the area of Question Answering system started in the year 1960 and at present lot of Question Answering systems have been developed. Question Answering system combines the research from different domains like Natural Language Processing, Artificial Intelligence, Information Retrieval and Information extraction. QA is made out of three particular modules. These three core segments are: question processing, document processing, and answer extraction. Question Processing plays an important role in QAS by classifying the submitted query according to its type. Information retrieval is important for question answering, because it find the answer relevant document from the corpora. Finally, answer extraction goal is to recover the response for a question posed by the user. In this paper we investigate various QAS. We give also statistics and analysis that can clear the way and help researchers to choose the appropriate solution to their issue. They can see the deficiency, so they can propose new systems for complex questions. They can likewise adjust or reuse QAS methods for specific research issues.**

keywords - **Question Answering System, Question Classification, Natural Language Question Answer, Information Retrieval , Natural Language Processing**

I. Introduction

The rapid increase in massive information storage and the popularity of using the Internet allow researchers to store data and make them available to the public. However, the exploration of this large amount of data makes finding information a complex and expensive task in terms of time. This difficulty has inspired the advancement of new reformed research tools, for example, Question Answering Systems.

Question Answering (QA) is a special field in the realm of Information Retrieval (IR). Alongside Information Retrieval it includes research from dissimilar, however related, fields which are Information Extraction (IE) and Natural Language Processing (NLP) and Machine Learning.

Current information retrieval system or search engine can do just “document retrieval”, i.e. given some keywords it only returns the relevant ranked documents that contain these keywords. IR systems don't return answers, and thus users are left to find answers from the documents by themselves. However, what a user really wants is often a precise answer to a question [1], [2]. Subsequently, the primary target of all QA systems is to recover answers to questions instead of full documents or most identical passages. Typically an automated QA system has three stages [3]: question processing, document processing and answer extraction. Figure 1 illustrates the basic architecture of a factoid QA system.

This paper is arranged as follows: in the next section an overview of QAS is given. In section 2 we primarily discuss Question Answering System major components. Section 3 provides a past history of work done in field of Question Answering. Section 4 analyze and discuss types of effective evaluation metrics and some most popular QAS based on their contribution, experimental results & limitations. In final section we summarize our survey work.

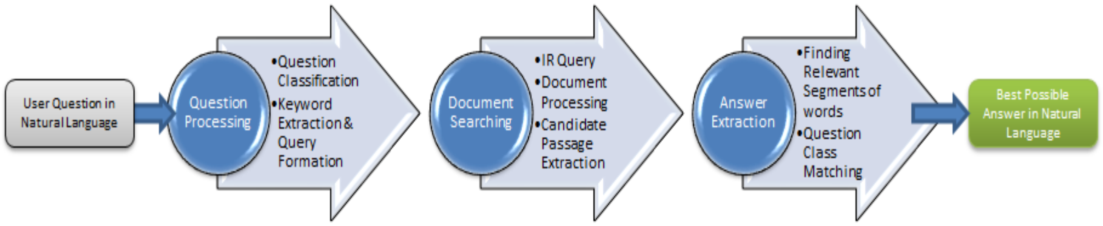

User Question in Natural Language
Question Processing
•Question Classification
•Keyword Extraction & Query Formation
Document Searching
•IR Query
•Document Processing
•Candidate Passage Extraction
Answer Extraction
•Finding Relevant Segments of words
•Question Class Matching
Best Possible Answer in Natural Language

Fig 1. QAS Basic Architecture

II. Question Answering System Components

As appeared in (Fig. 1), a usual QA system comprises of three discrete modules and each one of which has a primary component along with others. A typical QAS consist of three diverse modules, each of which has a core component alongside other supplementary components: “Query Processing Module” whose fundamental objective is QC (question classification), the “Document Processing Module” possess main function of information retrieval, and “Answer Extraction Module” has main work of answer finding.

Usually, the below scenario happens in the QA System: First, the user put a query on the QA system, the following scenario occurs in the QAS:

- First, the user posts a question to the QA system.
- Next the question analyser decides the focal point of the query to improve the exactness of the QA system.
- Question classification does a vital job in the QA system by recognizing the question type and hence the type of the desired answer.
- In question reformulation, the original question is phrased differently by expanding the question and passing it to the IR system.
- The IR component is utilized to recover the pertinent documents depending on significant keywords appearing in the question.
- The recovered pertinent documents are filtered and reduced into passages which are expected to contain the appropriate answer.
- These filtered passages are arranged and given to answer processing module.
- Based on the answer type and other recognition techniques, the candidate answers are identified.
- Some set of heuristics is created to extract only the related words or phrase which answers the question.
- The extracted answer is finally validated for its correctness and presented to the user.

Question processing module identifies focus point of the question, categories question type, find the anticipated answer type, and re-phrase the question into semantically equal multiple questions.

Re-formulation of a query into semantically alike questions is called query expansion and it grows the recall of the IR system. In Information retrieval system recall is quiet important for QA, because if none of the answer is correct in a document, no further processing can be done to find an answer [4].

2.1 Question Processing

The job of question processing is to break down the question and make an appropriate IR query and detect the entity type of answer, a category name which determines the type of answer. The first work is called query reformation and the later one is called question classification.

2.2 Document Processing

The task of document processing is to query over the IR engine, process the returned documents and return candidate passages that are likely to contain the appropriate response. Question classification helps here: it can decide the search technique to recover candidate passages. Depending on the question class, the search query can be converted into a form which is most suitable for finding the answer.

2.3 Answer Processing

The final task of a QA system is to process above obtained candidate passages after doing document processing and extract a segment of word(s) that is likely to be the answer of the question. Question classification again comes handy here. The candidate answers are ranked according to their likelihood of being in the same class as question class and the top ranked answer(s) will be considered as the final answer(s) of the question.

Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts. True Type 1 or Open Type fonts are required. Please embed all fonts, in particular symbol fonts, as well, for math, etc.

III. Related Work

The study to build a system which answers natural language questions backs to early 1960s. The first question answering system, baseball, [5] was able to answer domain-specific natural language questions which was about the baseball games played in American league over one season. This system was essentially a DB-centered system which used to convert a natural language question to a query on database. Most of the earlier studies [6] [7] were mainly domain-specific and had many limitation on answering questions. Because of absence of enough back-end learning to give answer to open domain questions, the question answering research laid dormant for few decades until the rise of the web. The enormous amount of information on the web on one hand and the requirement for questioning the web on other hand, brought again the work of question answering into lime light. After the TREC (Text REtrieval Conference) began QA track in 1998, focus on QA research again picked momentum [8].

Open-domain question answering system deals with all type of questions and can only rely on general ontology and information for instance WWW (World Wide Web). While, closed-domain QA deals with questions of a specific domain (music, sport, education etc.). The domain specific question answering system involves ample use of natural language processing systems supported by building domain centred ontology. [9]

The simplest type of question answering systems are dealing with factoid questions. The appropriate answer of this sort of questions are either one word or more which provide the exact answer of the question. For instance questions like “What was the name of Maharana Pratap’s horse?” or “Who discovered zero?” are factoid questions. Sometimes the question asks for a body of information instead of a fact. Instance of such questions are “What is Euphoria ?” or “Why did the world enter a global depression in 1929?”.

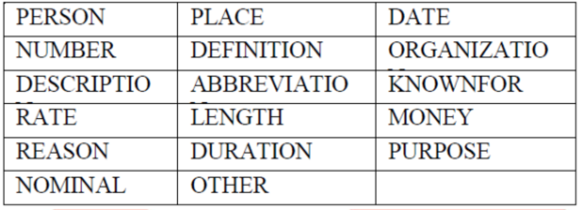
To answer these questions typically a summary of one or more documents should be given to the user. Many techniques from information retrieval, natural language processing and machine learning have been employed for question answering systems. Some early investigations were primarily based on querying structured information while the others used to apply pattern matching procedures. [10] gives an overview of the early QA systems. Ongoing investigations on open-domain question answering systems are normally founded on Information Retrieval (IR) methods. The QAS based on IR try to discover the answer of any specified question by treating a corpus of documents, usually from the web, and finding passage which is most likely to be the answer of given question. Some other recent works are founded on some pre-defined ontologies. These systems depend on semi-structured KB (knowledge bases) and cannot right away process natural language documents on the web. They often demand the web documents to be represented in structured or semi-structured formats. Semantic web [11] was the best endeavor to show the web documents in an organized manner; in spite of the fact that it never accomplished its ideal state. QA Systems, e.g. START [12], and True Knowledge [13] are two QAS backed by semi-structured information and semantic web-based techniques. These system possess their own knowledge bases (KB) which are primarily made by semi-automated data tagging.

In any way, QA systems evolution in past few decades reached till current stage which has significant improvement in providing answers in natural languages. QA systems, as mentioned before, have a backbone composed of three main parts: question classification, document processing, and answer extraction. Therefore, all three components attracted the focus of QA researchers.

3.1 Question Classification

Questions generally comes in predictable patterns and hence classified based on taxonomies. Taxonomies are distinguished into two main types: flat and hierarchical taxonomies. Flat taxonomies have one level of classes and no sub-classes, while hierarchical taxonomies possess multi-level classes. [6] suggested "QUALM", a software that uses a conceptual arrangement of 13 conceptual classes. [14] proposed QAS, NSIR (it is pronounced "answer"), which utilize a flat hierarchy with 17 classes, shown in (Table 1).

Table 1: Flat Taxonomy

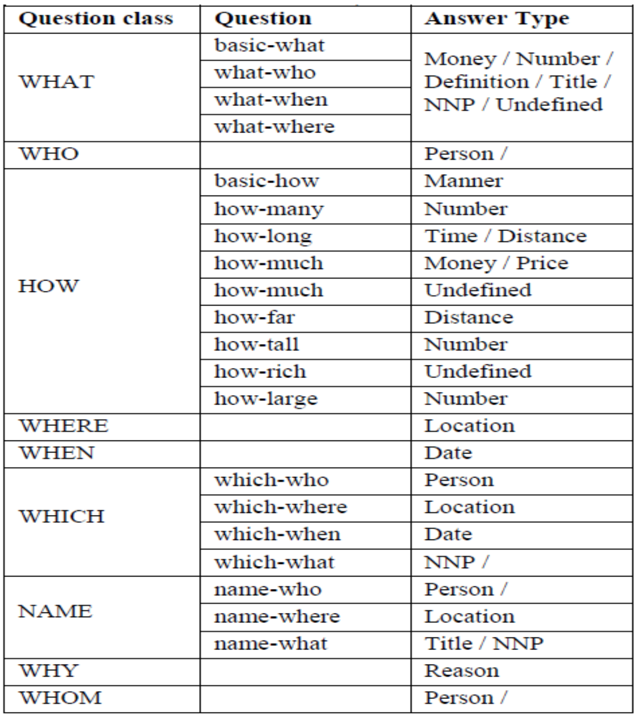


PERSON	PLACE	DATE
NUMBER	DEFINITION	ORGANIZATIO
DESCRIPTIO	ABBREVIATIO	KNOWNFOR
RATE	LENGTH	MONEY
REASON	DURATION	PURPOSE
NOMINAL	OTHER	

In the proceedings of TREC-8 [15], [16] offered a hierarchical nomenclature (Table 2) that classified the question types into nine classes, each of which was divided into a number of subclasses. These question classes and subclasses covered all 200 questions in the TREC-8 corpus.

[17] used a taxonomy having categories connected to several word classes of the WordNet ontology. Recently, in TREC-10 proceedings [15], [18] proposed a 2-layered classification, given in Table 3, which had six coarse grained and fifty fine grained classes. As a further step after setting the taxonomy, questions are classified based on that taxonomy using two main approaches: rule-based classifiers and machine learning classifiers.

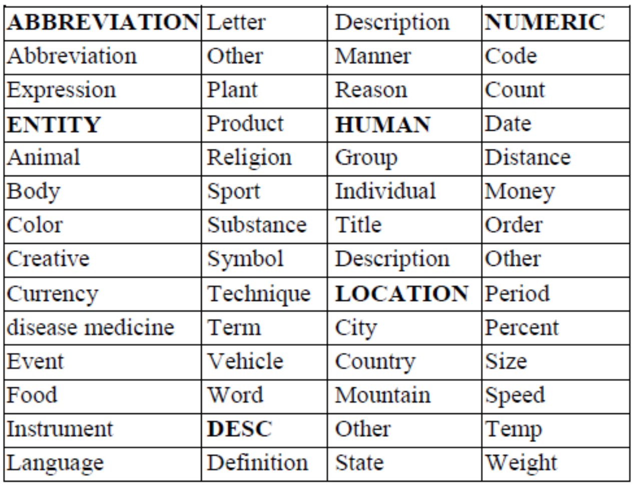
Table 2: Hierarchical Taxonomy



Question class	Question	Answer Type
WHAT	basic-what	Money / Number / Definition / Title / NNP / Undefined
	what-who	
	what-when	
	what-where	
WHO		Person /
HOW	basic-how	Manner
	how-many	Number
	how-long	Time / Distance
	how-much	Money / Price
	how-much	Undefined
	how-far	Distance
	how-tall	Number
	how-rich	Undefined
	how-large	Number
WHERE		Location
WHEN		Date
WHICH	which-who	Person
	which-where	Location
	which-when	Date
	which-what	NNP /
NAME	name-who	Person /
	name-where	Location
	name-what	Title / NNP
WHY		Reason
WHOM		Person /

Apparently, the rule-based classification is a direct way of classifying a question as per the taxonomy through a set of pre-defined heuristic rules. The rules could be just simple as, for example, the questions starting with "Where" are classified as of type LOCATION, etc. Many researchers used this approach due to its ease and quickness like [16], [15], as well as [19] who utilized both approaches, the rule-based and ML based.

Table 3: Hierarchical Taxonomy



ABBREVIATION	Letter	Description	**NUMERIC**
Abbreviation	Other	Manner	Code
Expression	Plant	Reason	Count
ENTITY	Product	**HUMAN**	Date
Animal	Religion	Group	Distance
Body	Sport	Individual	Money
Color	Substance	Title	Order
Creative	Symbol	Description	Other
Currency	Technique	**LOCATION**	Period
disease medicine	Term	City	Percent
Event	Vehicle	Country	Size
Food	Word	Mountain	Speed
Instrument	**DESC**	Other	Temp
Language	Definition	State	Weight

In machine learning approach, a machine learning model is designed and trained on an annotated corpus composed of labeled questions. The approach assumes that useful patterns for later classification will be automatically captured from the corpus. Therefore, in this approach, the choice of features (for representing questions) and classifiers (for automatically classifying questions into one or several classes of the taxonomy) are very important. Features could vary from basic surface of word or morphological ones to in depth syntactic and semantic features via linguistics analysis. [19] utilized ML based parsing and question classification for QA. [20] compared various choices for machine learning classifiers using the hierarchical taxonomy proposed by [18], for example: Support Vector Machines (SVM), Nearest Neighbors (NN), Naïve Bayes (NB), Decision Trees (DT), and Sparse Network of Winnows (SNoW).

3.2 Information Retrieval

[21] demonstrated a document retrieval investigation on a QAS, and assessed the usage of named entities (NE) and of noun, verb, and preposition phrases as accurate match in a document retrieval query. [21] defined an approach to question answering which was based on linking an information retrieval system with a natural language processing system that performed sensibly thorough linguistic analysis. While [22] introduced a basic approach to improve the precision of a QAS utilizing a knowledge database to directly get the similar answer for a question that was previously submitted to the QA system, and whose answer has been previously validated by the user.

3.3 Answer Extraction

[23] presented a model for finding answers by exploiting surface text information using manually constructed surface patterns. In order to enhance the poor recall of the manual hand-crafting patterns, many researchers acquired text patterns automatically such as [24]. Likewise, [25] showed an approach to deal with capturing long-distance dependence by utilizing linguistic structures to improvise patterns. Instead of exploiting surface text information using patterns, many other researchers such as [26] employed the named-entity approach to find an answer.

to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

IV. Analysis And Discussion

This section discusses and analyses the popular models proposed by QA researchers. Researches are showed and discussed in a sequential order describing the main contributions, results, and primary limitations for each research. However, as an initial subsection, metrics used in assessing QAS are first shown to give a detailed explanation of the meaning behind the experimental results obtained by the QA researches. At the end of discussion, a following small section reviews and concludes what had been studied and discussed.

4.1 Evaluation Metrics

The evaluation of QA systems is determined according to the criteria for judging an answer. The following list captures some possible criteria for answer evaluation [1]:

- Relevance: the answer should be a response to the question.
- Correctness: the answer should be factually correct.
- Correctness: the answer ought to be correct with respect to facts. Conciseness: the appropriate response should not contain unimportant or immaterial data.
- Completeness: the answer should be complete (not a part of the answer).
- Justification: the answer should be supplied with sufficient context to allow a user to determine why this was chosen as an answer to the question.

In light of the previously mentioned criteria, there are three distinct decisions for an answer extricated from a document:

- “Correct”: if the answer is receptive to a question in a right manner - (criteria 1 & 2)
- “Inexact”: if some data is missing from or added to the answer - (criteria 3 & 4)
- “Unsupported”: if the answer is not supported via other documents - (criterion 5)

There are many evaluation metrics which are different from one question answering campaign to another (e.g. TREC, NTCIR, CLEF, etc.). Moreover, some researchers develop and utilize their own customized metrics. However, the following measures are the most commonly used measures that are typically utilized for automated evaluation:

4.1.1 Precision, Recall and F-measure

Precision and recall are the conventional measures which have been used for long time in IR while F-measure is harmonic mean of precision and recall; these three metrics are given by:

$$Precision = \frac{count\ of\ correct\ answers}{count\ of\ question\ answered}$$

$$Recall = \frac{count\ of\ correct\ answers}{count\ of\ question\ to\ be\ answered}$$

$$F - Measur = \frac{2(Precsion\ x\ Recall)}{Precision + Recall}$$

4.1.2 Mean Reciprocal Rank (MRR)

Mean Reciprocal Rank (MRR), which (first utilized for TREC-8), is used to compute the answer rank (relevance):

$$MRR = \sum_{i=1}^{n} \frac{1}{\mathrm{r}_i}$$

where n is the total number of test questions and r_i is the rank of first accurate answer for the i-th test question.

4.1.3 Confidence Weighted Score (CWS)

The confidence about the accuracy of an answer is assessed utilizing another metric called Confidence Weighted Score (CWS), which was defined for TREC11:

$$\text{CWS} = \sum_{i=1}^{n} \frac{P_i}{n}$$

where n is total number of test questions and P_i is the accuracy of the answers at positions from 1 to i in the well-ordered list of answers.

4.2 Existing Research – Contributions, Experiments and Limitations

4.2.1 Moldovan et al. (LASSO) [16], 1999

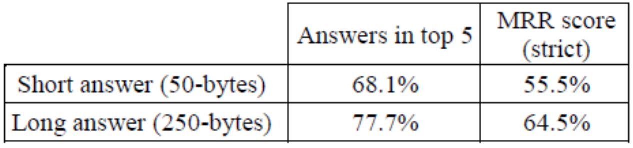
- **Contribution**

Their research count on NLP methodology in innovative ways to find answers in huge collections of documents. The question was processed by joining syntactic, semantic information that describe the question (e.g. question type or question focus), in which eight heuristic rules were defined to extract the keywords used for identifying the answer. The research also presented paragraph indexing where recovered documents were first filtered in paragraphs and then ordered.

- **Experimental environment and results**

The test condition was made out of 200 questions of the TREC-8 corpus, where all questions were categorized according to a hierarchy of Q-subclasses.

Table 4: Experimental Results – [16]



	Answers in top 5	MRR score (strict)
Short answer (50-bytes)	68.1%	55.5%
Long answer (250-bytes)	77.7%	64.5%

- **Limitations**

The question was considered to be answered correctly just if it was among the top five ranked long answers. In spite of the fact that, this was not viewed as an issue around then, yet beginning from TREC-2002, it was required for all QAS to give just one answer.

4.2.2 Harabagiu et al. (FALCON) [17], 2000

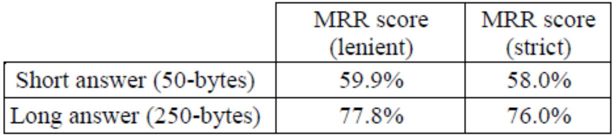
- **Contribution**

The same originators of LASSO [16] proceeded with their work and proposed another QAS called FALCON which used the same architecture of LASSO. The newly suggested system, FALCON, was described by additional features and components. They made a retrieval model for increasing knowledge in the answer engine through WordNet for semantic treatment of questions. Also, in order to get rid of the main limitation that found in LASSO, they gave a reasoning to rule-out inaccurate answers to give only one answer.

- **Experimental environment and results**

The experiments have been performed on TREC-9 corpus where questions and documents were more than that of TREC-8 and of a higher level of complexity. The test results of FALCON beaten those of LASSO, it proved that the added features had improved the preceding model.

Table 5: Experimental Results – [17]



	MRR score (lenient)	MRR score (strict)
Short answer (50-bytes)	59.9%	58.0%
Long answer (250-bytes)	77.8%	76.0%

4.2.3 Hermjakob [18], 2001

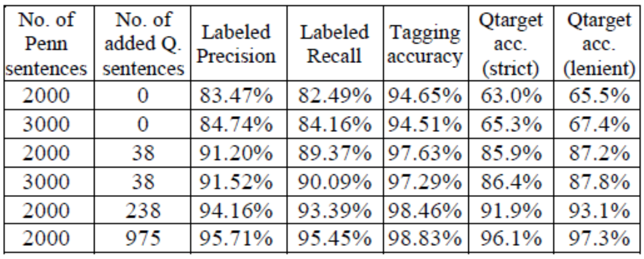
- **Contribution**

The research presented that parsing enhanced dramatically when Penn Treebank training corpus was augmented with an additional Questions Treebank, where parse trees were semantically enriched to assist QA matching. The research also defined the hierarchical structure of diverse answer types “Qtargets” where questions were classified.

- **Experimental environment and results**

In first two test runs, system was trained on 2000 and 3000 WSJ (Wall Street Journal) sentences (augmented Penn Treebank). In third and fourth runs, the parser got trained with same WSJ sentences empowered by 38 tree-banked pre-TREC-8 questions. For fifth run, 200 more questions added in TREC-8 as training sentences for testing TREC-9 sentences. In final run, the TREC-8 and TREC-9 questions were distributed into 5 subsets of about 179 questions. The system got trained on 2000 WSJ sentences and 975 questions.

Table 6: Experimental Results – [18]



No. of Penn sentences	No. of added Q. sentences	Labeled Precision	Labeled Recall	Tagging accuracy	Qtarget acc. (strict)	Qtarget acc. (lenient)
2000	0	83.47%	82.49%	94.65%	63.0%	65.5%
3000	0	84.74%	84.16%	94.51%	65.3%	67.4%
2000	38	91.20%	89.37%	97.63%	85.9%	87.2%
3000	38	91.52%	90.09%	97.29%	86.4%	87.8%
2000	238	94.16%	93.39%	98.46%	91.9%	93.1%
2000	975	95.71%	95.45%	98.83%	96.1%	97.3%

4.2.4 Radev et al. (NSIR) [15], 2002

- **Contribution**

They showed a probabilistic way for Web-based NLQA (Natural Language Question Answering), called probabilistic phrase re-ranking (PPR). Their NSIR system used a flat taxonomy of seventeen classes, where two methods were used to categorize the questions; the ML approach using a decision tree classification, and a experiential rule-based approach.

- **Experimental environment and results**

The system was assessed upon the 200 question from TREC-8, in which it attained a total reciprocal document rank of 0.20. The accuracy in classifying questions had been greatly improved using heuristics. Using machine learning, the training error rate was around 20% and the test error rate reached 30%. While the training error in the heuristic approach never exceeded 8% and the testing error was around 18%.

- **Limitations**

The PPR method didn't achieve the anticipated promising results due to simple sentence separation and POS tagging and text chunking. Also, their QA system did not reformulate the query submitted by the user.

4.2.5 Ravichandran & Hovy [23], 2002

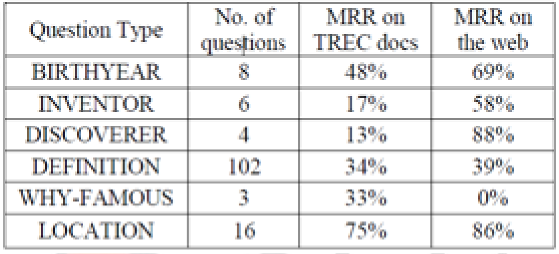
- **Contribution**

They presented a method that learns patterns from online data using some seed questions and answer anchors, without needing human annotation.

- **Experimental environment and results**

Using the TREC-10 question set, 2 set of experiments were performed. In first one, the TREC corpus was used as the input source using an IR component of their QA system. In the second experiment, the web was used as the input source using AltaVista search engine to perform IR.

Table 7: Experimental Results – [23]



Question Type	No. of questions	MRR on TREC docs	MRR on the web
BIRTHYEAR	8	48%	69%
INVENTOR	6	17%	58%
DISCOVERER	4	13%	88%
DEFINITION	102	34%	39%
WHY-FAMOUS	3	33%	0%
LOCATION	16	75%	86%

- **Limitations**

It only worked for certain types of questions that had fixed anchors, such as "where was X born". Therefore, it performed poorly with common definitional questions, since patterns didn't support long-distance dependencies.

4.2.6 Li & Roth [19], 2002

- **Contribution**

Their main contribution was proposing a hierarchical taxonomy in which questions were classified and answers were identified upon that taxonomy. [19] utilized and tested a ML technique called SNoW to classify the questions into coarse and fine grained classes of the taxonomy. They also showed through another experiment the differences between a hierarchical and flat classification of a question.

- **Experimental environment and results**

Their experiments utilized about 5500 questions distributed into 5 different sizes datasets, collected from 4 different sources. These datasets were used to train their classifier, which was then tested using 500 other questions collected from TREC10. Their experimental results demonstrated that the question classification (QC) problem can be solved quite precisely using a learning approach.

- **Limitations**

The research did not consider or test other machine learning classifiers that could have achieved more accurate results than SNoW, and at the same time it did not provide any reason for choosing SNoW in particular over other machine learning algorithms.

4.2.7 Zhang and Lee [20], 2003

- **Contribution**

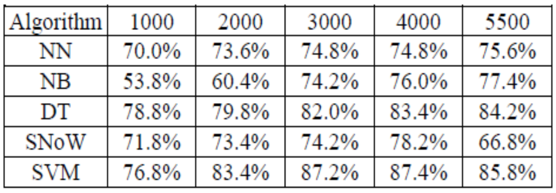
This research worked on the limitation of the aforementioned research [19], and carried out a comparison between five different algorithms of machine learning which were: Naïve Bayes (NB), Nearest Neighbors (NN), Support Vector Machine

(SVM), Decision Tree (DT) and Sparse Network of Winnows (SNoW). Additionally, they proposed special kernel function (tree kernel) that was computed proficiently by dynamic programming (DP) to empower the SVM to take advantage of the syntactic arrangements of questions which were helpful in question classification.

- **Experimental environment and results**

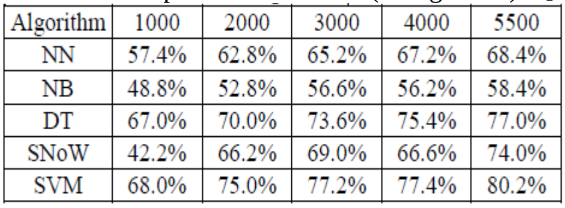
Under the same experimental environment used by [19], all learning algorithms were trained on 5 different sizes training datasets and then tested on TREC-10 questions. The experimental results demonstrated that the SVM algorithm outdone the four other methods in classifying questions either under the coarse category (Table 8), or under the fine category (Table 9). The question classification performance was measured by accuracy, i.e. the proportion of correctly classified questions among all test questions.

Table 8: Experimental Results (coarse-grained) – [20]



Algorithm	1000	2000	3000	4000	5500
NN	70.0%	73.6%	74.8%	74.8%	75.6%
NB	53.8%	60.4%	74.2%	76.0%	77.4%
DT	78.8%	79.8%	82.0%	83.4%	84.2%
SNoW	71.8%	73.4%	74.2%	78.2%	66.8%
SVM	76.8%	83.4%	87.2%	87.4%	85.8%

Table 9: Experimental Results (fine-grained) – [20]



Algorithm	1000	2000	3000	4000	5500
NN	57.4%	62.8%	65.2%	67.2%	68.4%
NB	48.8%	52.8%	56.6%	56.2%	58.4%
DT	67.0%	70.0%	73.6%	75.4%	77.0%
SNoW	42.2%	66.2%	69.0%	66.6%	74.0%
SVM	68.0%	75.0%	77.2%	77.4%	80.2%

4.2.8 Peng et al. [25], 2005

- **Contribution**

Their research presented an approach to handle the main limitations of [23]. They discovered a hybrid method for Chinese definitional QA by combining deep linguistic analysis (e.g. co-reference, parsing, named-entity) and shallow pattern learning to capture long-distance reliance in definitional questions.

- **Experimental environment and results**

They created a list of questions and recognized answer snippets from TDT4 data. The overall results exhibited that combining both pure linguistic and pure pattern-based systems enhanced the performance of definitional questions, which proved that linguistic analysis and pattern learning were complementary to each other, and both were helpful for definitional questions.

- **Limitations**

The pattern matching was based on simple Part-of-Speech (POS tagging which took only limited syntactic information without showing any semantic information.

4.2.9 Stoyanchev et al. (StoQA) [4], 2008

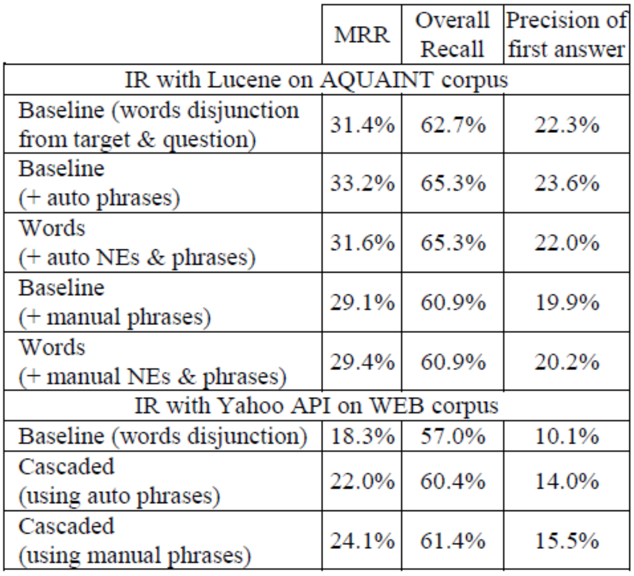
- **Contribution**

In their study, they demonstrated a document retrieval experiment on a QAS. constituents to look through search queries. The procedure of extracting phrases was done with the help of named-entity recognition (NER), parts-of-speech taggers and stop-word lists.

- **Experimental environment and results**

The QAS was assessed using two datasets: the AQUAINT corpus, a 3GB collection of news documents utilized in TREC-2006; and the other dataset was the Internet. The datasets utilized TREC questions with non-blank answers. The documents in AQUAINT corpus indexed using Lucene engine. Their experiments utilized automatically and manually created phrases. The automatically formed phrases were found by extracting nouns, verbs, and propositional phrases, while the manually made phrases were received by hand-correcting these automatic annotations.

Table 10: Experimental Results – [4]



	MRR	Overall Recall	Precision of first answer
IR with Lucene on AQUAINT corpus			
Baseline (words disjunction from target & question)	31.4%	62.7%	22.3%
Baseline (+ auto phrases)	33.2%	65.3%	23.6%
Words (+ auto NEs & phrases)	31.6%	65.3%	22.0%
Baseline (+ manual phrases)	29.1%	60.9%	19.9%
Words (+ manual NEs & phrases)	29.4%	60.9%	20.2%
IR with Yahoo API on WEB corpus			
Baseline (words disjunction)	18.3%	57.0%	10.1%
Cascaded (using auto phrases)	22.0%	60.4%	14.0%
Cascaded (using manual phrases)	24.1%	61.4%	15.5%

- **Limitations**

The experimental results showed that the overall accuracy on the web was lower than that on the AQUAINT corpus.

4.2.10 Kangavari et al. [21], 2008

- **Contribution**

Their exploration gave a model to improve QAS by using query reformulation and validating the answer. The model depends on previously asked questions along with the user feedback (voting) to reformulate questions and validate answers through the domain knowledge database.

- **Experimental environment and results**

The system was functioning on a closed aero-logic field for predicting weather information. Results presented that, from a total 50 asked questions, the model attained 92% improvement.

- **Limitations**

The model was tested in a restricted experimental environment in which the domain was very specific and the number of questions is relatively small. Also, relying only on the users as a single source for authenticating answers is a double-edged weapon.

4.3 Conclusion of Analysis & Discussion

The period within (1999-2007) was very rich in QA research than any other time. This was mostly because of the interesting research environment provided by QA tracks of the TREC annual conferences of that period. Yet, other campaigns such as NTCIR and CLEF still represent an important nerve for QA research.

Likewise, most researches of QA field were somehow diverse with respect to their system design, methodologies, scope, evaluation metrics, etc. But, on the other hand researches were mainly concerned with one or more of the three basic components of QA systems: question classification, document processing and answer extraction, which combine methods from NLP (natural language processing), IR (information retrieval), and IE (information extraction) respectively.

V. Survey Conclusion

This survey paper organized and summarized recent researches in a novel and cohesive manner that added understanding to work in the QA field. It underlined the classification of the existing literature, developing a viewpoint on the area, and evaluating trends.

However, because it is difficult for a survey to comprise all or even most of earlier research, this survey included only the work of the top-published and cited authors of the question answering (QA) field. Furthermore, because scientific research is a progressive, continuously on-going and accumulative endeavour , this survey also included research having minor limitations to illustration how these limitations were revealed, encountered and treated by other researchers.

VI. References

[1] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," Natural Language Engineering, vol. 7, no. 4, pp. 275-300, 2001.

[2] D. Zhang and W. Lee, "A Web-based Question Answering System," Massachusetts Institute of Technology (DSpace@MIT), 2003

[3] Daniel Jurafsky and James H Martin, “Speech and Language Processing”, (2nd Edition) (Prentice Hall Series in Artificial Intelligence) Prentice Hall, 2008

[4] S. Stoyanchev, Y. Song and W. Lahti, "Exact phrases in information retrieval for question answering," in Proceedings of the 2nd workshop on Information Retrieval for Question Answering, 2008.

[5] Green, B.f., Chomsky, C., Laughery, K., 1961. BASEBALL, “An automatic question answerer”, Proceedings of the Western Joint Computer Conference, New York: Institute of Radio Engineers, pp. 219–224

[6] W. G. Lehnert, The Process of Question Answering - A Computer Simulation of Cognition, Yale University, 1977.

[7] W. Woods, "Progress in Natural Language Understanding: An Application to Lunar Geology," in Proceedings of the National Conference of the American Federation of Information Processing Societies, 1973.

[8] David A. Hull. Xerox, “TREC-8 question answering track report”, In Voorhees and Harman, 1999

[9] M. Ramprasath and S. Hariharan, "A Survey on Question Answering System," International Journal of Research and Reviews in Information Sciences (IJRRIS), pp. 171-179, 2012.

[10] Androutsopoulos I., Ritchie G.D. and Thanisch P., “Natural Language Interfaces to Databases – An Introduction. Natural Language Engineering”. 1(1): 29-81.Cambridge University Press. (1995)

[11] Berners-Lee T., Hendler J. and Lasilla O., “The Semantic Web”. Scientific American. (2001)

[12] Katz, Felshin, B., Yuret, S., Ibrahim, D., Temelkuran, B., 2002, “Omnibase: uniform access to heterogeneous data for question answering” Proc of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB)

[13] Tunstall-Pedoe, William. (2010). True Knowledge: Open-Domain Question Answering Using Structured Knowledge and Inference. AI Magazine. 31. 80-92. 10.1609/aimag.v31i3.2298.

[14] D. Radev, W. Fan, H. Qi, H. Wu and A. 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.

[15] E. Voorhees, "Overview of the TREC 2002 Question Answering Track," in Proceedings of the Text Retrieval Conference (TREC 2002), 2002.

[16] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju and V. Rus, "Lasso: A Tool for Surfing the Answer Net," in Proceedings of the Eighth Text Retrieval Conference (TREC-8), 1999.

[17] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus and P. Morarescu, "FALCON: Boosting Knowledge for Answer Engines," in Proceedings of the Ninth Text Retrieval Conference (TREC9), 2000.

[18] U. Hermjakob, "Parsing and Question Classification for Question Answering," in Proceedings of the Workshop on Open-Domain Question Answering at ACL-2001, 2001.

[19] X. Li and D. Roth, "Learning question classifiers," in Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), 2002.

[20] D. Zhang and W. Lee, "Question Classification using Support Vector Machines," in Proceedings of the 26th Annual International ACM SIGIR Conference, 2003.

[21] 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.

[22] M. Kangavari, S. Ghandchi and M. Golpour, "Information Retrieval: Improving Question Answering Systems by Query Reformulation and Answer Validation," World Academy of Science, Engineering and Technology, pp. 303-310, 2008.

[23] D. Ravichandran and E. Hovy, "Learning Surface Text Patterns for a Question Answering System," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002.

[24] J. Xu, A. Licuanan and R. Weischedel, "TREC2003 QA at BBN: Answering Definitional Questions," in Proceedings of the 12th Text Retrieval Conference, 2003.

[25] F. Peng, R. Weischedel, A. Licuanan and J. Xu, "Combining deep linguistics analysis and surface pattern learning: A hybrid approach to Chinese definitional question answering," in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, 2005.

[26] C. Lee, C. Shih, M. Day, T. Tsai, T. Jiang, C. Wu, C. Sung, Y. Chen, S. Wu and W. Hsu, "ASQA: Academia Sinica Question Answering System for NTCIR-5 CLQA," in Proceedings of NTCIR-5 Workshop Meeting, Tokyo, 200