Journal of Theoretical and Applied Information Technology
31st July 2017. Vol.95. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

QUESTION ANSWERING SYSTEM : A REVIEW ON QUESTION ANALYSIS, DOCUMENT PROCESSING, AND ANSWER EXTRACTION TECHNIQUES

[1,2]FANDY SETYO UTOMO, [3]NANNA SURYANA, [4]MOHD SANUSI AZMI

[1]Department of Information Systems, STMIK AMIKOM Purwokerto, Purwokerto, Indonesia
[2,3,4]Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
E-mail: [1]fandy_setyo_utomo@amikompurwokerto.ac.id, [3]nsuryana@utem.edu.my, [4]sanusi@utem.edu.my

ABSTRACT

Question Answering System could automatically provide an answer to a question posed by human in natural languages. This system consists of question analysis, document processing, and answer extraction module. Question Analysis module has task to translate query into a form that can be processed by document processing module. Document processing is a technique for identifying candidate documents, containing answer relevant to the user query. Furthermore, answer extraction module receives the set of passages from document processing module, then determine the best answers to user. Challenge to optimize Question Answering framework is to increase the performance of all modules in the framework. The performance of all modules that has not been optimized has led to the less accurate answer from question answering systems. Based on this issues, the objective of this study is to review the current state of question analysis, document processing, and answer extraction techniques. Result from this study reveals the potential research issues, namely morphology analysis, question classification, and term weighting algorithm for question classification.

Keywords: *Information Retrieval, Question Answering, Question Analysis, Natural Language Processing.*

1. INTRODUCTION

Information retrieval (IR) approach on the traditional search engines like Yahoo, Google, and Bing uses keywords entered by the user. The search engine then provides information, according to the given keyword by users. In some cases, the information provided is suitable for the needs of users, but often the information provided is not relevant. The proposed Question Answering systems (QAS) has been considered to be able to solve these problems by providing an interface, where users can express their need for information in the form of Natural Language (NL) and the search engine will provide relevant answers to these questions [1–3]. Expression of information needs in Natural Language can be in the form of questions or statements [4]. In other words, QAS is a technology used to locate, extract, and give an accurate answer to the users question in the form of natural language [5–9]. According to [10], output from QAS isn't a list of documents, but more specific answers. Confirmed by [11] and [12], output QAS is a direct answer to the question, not a list of references that have possible answers, so that users have a minimum number of reading.

Question Answering System of the first generation appeared before 1965, had a limitation on linguistic model [13]. The second generation of QAS preceded by ELIZA computer program was built by [14] to learn communication between human and machine using Natural Language. Then, ELIZA was developed by Taylor in 1968 for computer aided instruction problem and counseling behavior simulation [15]. Research conducted by [16] developed ELIZA which was able to identify patterns of words entered by the user. The computer program they developed, able to answer "why" questions and "yes-no" questions. Research conducted by [17], developed conceptual parser system from a program which had been made previously by [16]. The system they created was the first QAS based on semantics.

Question Answering application has been applied in various fields to solve problems related to information retrieval on specific cases. Some of the areas that have implemented QAS such as social media ([10],[18],[19]), geographic ([5],[20–23]), geology [24], software engineering ([25],[26]),

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

aviation [27], biomedical ([28–30]), physics [7], biological [31], e-commerce [32], and religion ([33–35]). The wide use of Question Answering Systems in many fields indicates that QAS contributes to improving quality of life.

Question Answering Systems consists of several modules. Researches conducted by [1], [4], [6], [7], [11], [12], [18], [26], [28], [29], [36–42] have developed QAS using question analysis, document retrieval, and answer extraction module. However, some researchers such as [3], [5], [8], [21], [31], [43] have replaced the document retrieval on the development of QAS with the passage retrieval module. Researches conducted by [30], [44], [45] developed QAS using four modules, i.e. question analysis, document retrieval, passage retrieval, and answer extraction module.

Question Processing module transformed the user's questions in the form of natural language into phrases [41],[46],[47]. More specific has been described by [38], [48], [49] in their study, Question Processing consisted of two steps. First step, identify the semantic type of the entity according to the user question. Second step, specification additional restrictions on the answers entities, such as identifying keywords that will be used for searching answers and identifying relationships, syntactic, or semantic that must hold between a candidate answer entity and other entities. Once the user query is processed on the Question Processing Module, then the processed data is processed on Document Processing module. Document processing module consists of two parts, i.e. document retrieval and passage retrieval. Document retrieval with query reformulation technique is a process for identification candidate documents, which contains the answer that relevant to the user query [12][36][38][44]. Query reformulation is a keyword transformation technique within the question to another form using synonyms or semantic [1][47]. Once the candidate documents were obtained from the document retrieval process, passage retrieval component extracted small text passages or textual units that contained the answers [3][43][44][46][50]. Answer extraction component is the last stage of QAS. This component receives the set of passages from the previous module, then determine the best answers to user questions. Answer extraction consists of several stages, i.e. candidate answer extraction, answer scoring and ranking, and answer selection [3][43][44][51].

Challenge to optimize Question Answering Framework is to increase the performance of all modules in the framework. The performance of all modules that has not been optimized has led to the less accurate answer from QAS. Based on this issue, the objective of this study is to review the current state of question analysis, document processing, and answer extraction techniques.

This paper is organized as follows. Section 2 describe Question Answering Systems dimension classification. Section 3 presents the review of recent studies on question answering systems. Finally, Section 4 discuss open research issues.

2. QAS DIMENSION CLASSIFICATION

Question Answering System could be classified into several dimensions. Opinion among experts regarding dimensions classification can be differentiated between one another. According to [36], QAS was divided into three dimensions, i.e. answer source, domain coverage, and question analysis method. The first dimension observed the answers source came from unstructured or structured documents. The second dimension observed from the application domain, i.e. open domain or restricted/ closed domain. Open domain means QAS would extract answers from enormous data sources because it has a question that not only covers certain areas only. While closed or restricted domain has questions and answers from the data source for specific areas only. The third dimension is observed from the complexity of question analysis methods, such as used shallow/deep natural language processing, statistical methods, or semantic methods. Rule based pattern matching [39] and Hybrid approach [42] were other techniques that can be used to process user queries.

The detail of QAS dimension classification was explained by [44]. They classified into 5 dimensions, i.e. domain coverage, question analysis method, answer source, knowledge base offline/online, and the number of languages supported. A survey by [42] appended several dimensions from study [44], i.e. types of questions, types of matching functions used in document processing, and forms of answer generated by QAS.

Factoid and nonfactoid are the type of questions being asked by users in the QAS [42][44]. In addition, [42] added the type of questions, i.e. list type, hypothetical type questions, and confirmation questions. The expected answers depends upon the types of the questions asked by the users. Systems deals with different types of questions required different strategies to locate answers. Factoid questions consisted of *What*, *Where*, *Who*, *When*, *Which*, and *How much/many* [52–55]. Answers of factoid questions are date,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

time, duration, location, person, and organization entity. Whereas, non factoid consist of *How* and *Why* questions [56–59]. A *Why* question type, asked for reasons or causation, and a *How* question type, asked for a manner approach. Unlike factoid questions whose answer comprises a short text, nonfactoid questions were inherently more complex and require a paragraph-length answer [60]. Furthermore, types of matching functions used in document processing are Set-theoretic models, Algebraic models, Probability models, Feature-based models, and Conceptual graph based models [42]. Then, forms of answer generated by QAS are Extracted and Generated answer.

3. LITERATURE REVIEW

The research questions (RQ) were specified to keep the review focused. The research questions addressed by this literature review are:

- **RQ 1:**
  What are the essential components involved in Question Answering Systems?
- **RQ 2 :**
  How do the existing approaches perform Question Answering Systems?
- **RQ 3 :**
  How the recent approaches provides relevant information to the users?

A question answering system requires understanding of natural language processing (NLP) and linguistics techniques such as lexicon, tokenization, POS tagging, and parsing are implemented to user's question for formulating it into a precise query that merely extracts the respective response from the database [39][61]. NLP is a computer field and technique which is developed from language study and computational linguistic in artificial intelligence [12]. This technique was used to alter the question to a query that can be executed by the database, extract information from text, and retrieve relevant documents from a collection [62]. Several researchers have conducted previous studies for QAS development. Their studies involved the use of Natural Language Processing for Question analysis, document processing, or answer extraction. Table 1 describes some previous studies on QAS development.

Table 1: Previous Studies on QAS Development

Citation	Language	Input	Domain
[61],[63],[64], [65],[66],[51], [67]	English	Factoid	Restricted
[68],[69],[21], [28]	English	Factoid, statement	Restricted
[70],[12],[71], [72]	English	Factoid, nonfactoid	Restricted
[73],[74],[75], [76],[77],[78]	English	Factoid	Open
[79]	English	Nonfactoid, statement	Restricted
[80]	English	Factoid	Restricted, open
[81]	English	Factoid, confirmation	Open
[1]	English, arabic	Factoid, nonfactoid	Restricted
[82]	Malayalam	Factoid	Restricted
[6]	Hindi	Factoid	Restricted
[83]	Arabic	Statement	Restricted
[33]	English	Statement	Restricted
[84]	English	Nonfactoid	Restricted
[85]	Japanese	-	Open
[41],[86]	English	Factoid, nonfactoid	Open
[87]	Spanish	Factoid	Restricted
[35]	Arabic	Factoid	Restricted
[88]	Indonesian	Factoid	Restricted
[36]	English	Factoid, confirmation	Restricted
[7]	English	Factoid, confirmation, statement	Restricted
[89]	Arabic	Factoid, nonfactoid, confirmation	Open

Table 1 shows that there are several studies on QAS development which have different characteristics. Based on the complexity of question analysis methods, we classify the previous studies into several approaches, i.e., linguistic, statistical, semantic, rule-based pattern matching, and the hybrid approach that can be used to process user queries. Table 2 describes classification results of previous studies into several approaches.

Table 2: Classification Results of Previous Studies into Question Analysis Methods

Approach	Citation
Linguistic	[61],[68],[70],[63],[73],[79],[80],[81], [1],[69],[12],[82],[6],[83],[33].
Statistical	[84],[74],[85],[71],[41].
Semantic	[64],[72],[87],[75].
Rule-based Pattern Matching	[65],[66].
Hybrid	[86],[21],[51],[76],[77],[78],[35],[88], [28],[36],[67],[7],[89].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Based on Table 2, there are several studies using natural language processing approach to process user queries. Table 3 describes techniques and output from question analysis processing for each study using linguistic approach.

Table 3: Analysis Review of Question Processing Using Linguistic Approach

Cit.	Techniques	Output
[61]	Lexical analyzer, POS tagging	Query from inverse transformational grammar
[68]	Lexical analyzer, POS tagging	Parsing of a sentence
[70]	Named Entity Recognizer	Quasi logical form
[63]	Named Entity Recognizer	Keyword expansion
[73]	Language-to-logic, question focus recognition	Logical query representation (LQR)
[79]	Morphology operation and synonyms finder	Minimal Logical Forms (MLFs)
[80]	LinkParser [90]	Query formulation
[81]	Tokenizer, sentence splitter, POS tagger, and chunker provided by GATE [91]	Query triples (subject, predicate, object)
[1]	Spelling correction, query completion, stop words removal, diacritics removal operation, question classification, and query reformulation.	Query triples (subject, predicate, object)
[69]	Pre-lexicon processor, disambiguator, post-lexicon processor	SQL Query object
[12]	Tokenization, question checking, stopword removal, POS tagging, and stemming	Question keyword
[82]	Tokenization, stop word removal, Sandhi splitter, POS Tagging, and Vibhakthi Analysis	Keywords and question type
[6]	Tokenization, split sentence, and stop word removal	Keywords
[83]	Generate synonyms and semantic tags using Named Entities (NE)	Keywords with synonyms and semantic tags
[33]	WordNet and collection of islamic synonyms	Query expansion

Based on Table 3, most researchers used tokenization, lexical analyzer, morphology operation, POS tagging, Named Entity Recognizer, question classification, stopword removal, and WordNet to analyze user queries. Meanwhile, for the output of question analysis stage, they used keywords, query triples (subject, predicate, object), and query expansion.

According to [92], [93], [94], and [95] Question Classification is a vital component of any Question Answering system. Knowing the answer type would reduce processing and effort and provide a feasible way to select correct answers among the possible answer candidates. Named Entity Recognizer has a shortcomings, if the answer is a common noun, then the questions cannot be mapped with a named entity [75]. Solution for this issues is to use Semantic Role Labeling (SRL). For each predicate in a sentence, semantic roles identify all constituents, determining their roles and also their adjuncts.

The advantage using linguistics approach to process user queries is this technique has capability to provide a situation-specific answer [39][68]. However, this technique has drawbacks, i.e. different application domain requires different grammar and mapping rules. Additionally, building an appropriate knowledge base is a time-consuming process because knowledge base is very complex models [39][78][79]. Suffering from an opacity of linguistic and conceptual/ contextual coverage is another drawback from this technique [96].

Statistical approach is another technique to process user queries. This approache tries to exploit large amount of data to overcome the complex and time consuming tasks of pattern matching and information extraction. Statistical approach implements different statistical techniques, such as similarity computation, probability of relevance, mining, and filtering of N-grams to analyze questions for making prediction about users expected answer type [39][44][64]. Based on Table 2, there are five studies use statistical approach to process user queries. Table 4 describes techniques and output from question analysis processing for each study using statistical approach.

Table 4: Analysis Review of Question Processing Using Statistical Approach

Cit.	Techniques	Output
[84]	Question classification (using Maximum Entropy and applied n-grams)	Query expansion with named entity
[74]	Adaptive TF-IDF for term weighting	Query expansion
[85]	Question classification (using Support Vector Machine)	Query expansion
[71]	Named Entity Recognizer, question classification (using Rocchio and Support Vector Machine)	Term expansion, named entity, keywords
[41]	Statistical chunker (using Djikstra-style dynamic programming algorithm) to defined keyword	Keywords

Based on Table 4, most researchers used question classification to analyze user queries. Meanwhile,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

for the output of question analysis stage, they used query expansion, named entity, and keywords.

The advantage using statistical approach to process users queries is this technique could deal with large amount of data and their heterogeneity as well. However, this technique has drawbacks, i.e. it deal with each term independently and fail to identify linguistic features for combination of words [39][94].

Another technique to process user queries is using semantic analysis approach. This technique is able to recognize the possible meanings of the questions from words that used in question [42]. Based on Table 2, there are four studies used semantic approach to process user queries. Table 5 describes techniques and output from question analysis processing for each study usee semantic approach.

Table 5: Analysis Review of Question Processing Using Semantic Approach

Cit.	Techniques	Output
[64]	Semantic analysis to disambiguate syntactic analysis, question type and formalized the content of a question using ontology	Formal query TMR (Text Meaning Representation)
[72]	Question type, super concepts of nouns using WordNet, and determining semantic pattern	Query of semantic pattern
[87]	Applied fuzzy matching algorithm between the query and the ontology lexicon for word misspelt in user query	Query with Named entity and tagging
[75]	semantic roles using patterns, WordNet semantic classes to the patterns for filtering the potential answers	Keywords, question type, roles identification and adjuncts for each predicate in a query

Based on Table 5, most researchers used question type identification, semantic analysis and pattern, and WordNet to analyze user queries. Meanwhile, for the output of question analysis stage, they used keywords and specific query format.

Semantic analysis approach supports for term definition and query expansion processing to process users queries [95][96]. However, this technique has drawbacks, i.e. assuming that the source texts are expressed in natural sentences [26] and ontologies-semantic only cover a particular domain of knowledge [78].

Futhermore, Rule based pattern matching is another approach that can be used to analyze users query. Linguistic resources such as POS, Named Entity Recognizer, dictionaries, and WordNet, might be used to support process in a rule-based QAS [44]. Based on Table 2, there are two studies used this approach to process user queries. Research by [65] used several stages to analyze query, i.e. question classification, identifying primary and secondary terms using query expansion technique, detecting pattern, and ranking the snippets. Meanwhile, for the output of question analysis stage, they used primary and secondary terms (synonims). Study by [66] extended research [65] with complex pattern case. They added several steps to analyze query, i.e. checking for correlation between the pattern and the questions semantics, and identifying the exact answer in the complex pattern-matching string. For the output of question analysis stage, they used question classification and query expansion.

Rule based pattern matching approach is quite favourable for small and medium-size systems [39]. However, this technique has drawbacks, i.e. template-driven approaches have limitations because they cannot handle the variant in particular domain [30].

A hybrid approach is a combination between linguistic, statistical, semantic, and rule based pattern matching technique to analyze query. Based on Table 2, there are several studies used hybrid approach to process user queries. Table 6 describes techniques and output from question analysis processing for each study that used hybrid approach.

Table 6: Analysis Review of Question Processing Using Hybrid Approach

Cit.	Techniques	Output
[86]	Question features (using bag-of-words, bigrams, and trigrams) and question classification (using SVM and Sparse Network of Winnows)	Question classification
[21]	Question information extractor (using tokenization, POS tagging, Porter stemmer, and OAK Standford for named entity tagger) and question classification using rule based pattern matching from [100]	Named entity, question classification, and keywords
[51]	Tokenizing, POS tagging, Name Entity Recognition (NER), question classification using Maximum Entropy Classifier.	A weighted list of relevant keywords
[76]	POS tagging, parsing using lexical entries and pre-defined domain-independent lexical entries, semantic representation of the natural language query	SPARQL query template
[77]	Question classification (using SVM), question type (using Directed Acyclic Graph-SVM), tokenizing, POS tagging, and keywords extraction	Question classification, question type, and keywords

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[78]	POS Tagger, stopword removal, noun phrase pattern using entity recognition pattern	Entities based on entity recognition
[35]	Question pre-processing (Morphological analysis using POS tagger, diacritization, lemmatization, disambiguation, stemming, stop word removal. Remove pronouns, prepositions, conjunctions) and question classification using SVM.	Keywords and question classification
[88]	POS tagging, question type, and rule-based methods uses lexical and semantic heuristics	Question type, keywords, and keywords entity
[28]	Question pre-processing (using Tokenization and stop word removal), entity recognition (using semi-supervised learning, bootstrapping technique), and entity to a specific ontology class, property or instance.	Ontology class, property, and instance.
[36]	Question type (using pattern matching), question classification (using matching the NL questions with manually built lexical patterns), Medical Entity Recognition based on the new form of the question using A rule based method using the MetaMap tool and A statistical method using a CRF classifier, extraction of semantic relations.	One or several SPARQL queries (if have more than one question Type)
[67]	Identifying and extracting terms (Tokenizer, POS tagger, vector space model), terms expansion using WordNet (Adopt from [101]).	User's question with additional terms
[7]	Tokenization, stemming, POS tagging, stop word removal, and ontology lexicon. The ontology lexicon includes entity name, relation name and properties.	User's query is labelled with the ontology concepts.
[89]	Tokenization, POS tagger, identify interrogative noun using Rule for interrogative noun, question type, identify question scope according to pattern	Question noun, question type, question pattern, and scope.

Based on Table 6, most researchers used natural language processing on question processing stage. They used tokenization, POS tagging, stop word removal, and Stemmer for pre-processing. Furthermore, they used the results from pre-processing stage to determine question type, question classification, and entity recognition. For question classification, they used SVM, Sparse Network of Winnows, rule-based pattern matching, maximum entropy classifier, and matching the NL questions with manually built lexical patterns. Then, for question type recognition, they used SVM and pattern matching.

The results from question analysis stage is used as input on document processing stages. Natural language processing, semantic, statistical, rule-based, and hybrid techniques are the approach that could be used to retrieve documents or passages from the database. Table 7 describes classification results of previous studies on Table 1 into several approaches.

Table 7: Classification Results of Previous Studies into Document Processing Techniques

Approach	Citation
Linguistic	[73],[12],[82],[65],[88]
Statistical	[6],[33],[84],[74],[85],[71],[35],[89]
Semantic	[63],[81],[1],[64],[72],[87],[77],[28],[36]
Rule-based	[61],[68],[70],[79],[69],[66]
Hybrid	[83],[21],[51],[76] ,[7]
Other	[80],[41],[75],[86],[78]

Based on Table 7, there are several studies used linguistic approach to process the output from question analysis stage. Output from this stage becomes the input to the document processing stage. Table 8 describes document processing for each study that used linguistic approach and the answer extraction techniques. Another approach applied document processing techniques using Google, MSN, ASK, Altavista, and Gigablast search engine.

Table 8: Analysis Review of Document Processing using Linguistic Approach and Answer Extraction Techniques

Cit.	Document Processing	Answer Extraction
[73]	Paragraph finding based on keywords occurrences in the paragraphs. **Output** : Tagged paragraphs	**Entity based approach**. All paragraphs were weighted, and the top one was selected for answer
[12]	Keyword-matching technique. **Output** : tf (term frequency) for each document	Vector space model
[82]	Keyword-matching technique. **Output** : Sentences with weight of terms	Sentence with the highest weight of significant terms is selected as answer
[65]	Extracts query-related terms from the short relevant text passages. **Output** : Create snippets from the relevant text passages	Detecting patterns from snippets. Then, scoring and rank the snippets.
[88]	Keyword-matching technique and word match scoring function is applied to count number of similar words between question and document.	- Relevant documents are getting processed by rule-based scoring component to get final score. - Find the correct

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	Output : Relevant score for each document	answer within the highest scored document

Based on Table 8, most researchers used keyword-matching technique for retrieve information on document processing stage. Meanwhile, for answer extraction, most of them used the document's score and rank. The best answer is selected from the highest score.

Table 9 describes document processing for each study used statistical approach and the answer extraction techniques.

Table 9: Analysis Review of Document Processing using Statistical Approach and Answer Extraction Techniques

Cit.	Document Processing	Answer Extraction
[6]	1. Measure the similarity between the two texts using Cosine, Dice, Euclidean distance, Jaccard, JaroWinkler, SmithWaterman, Levenshtein, MongeElkan, and NeedlemanWunch similarity algorithm. 2. Determine question type using feature vector creation. 3. Classification method using the NN or KNN algorithm. **Output** : data training.	Answer prediction method using the Knowledge base
[33]	Artificial Neural Network to classify the verses of Al-Baqarah Surah. **Output** : Relevant verses from the Holy Quran.	1. Extract the answers using N-gram technique. 2. Words Matching scoring function to determine the best answer
[84]	1. Create expanded queries using the highest scoring passages. 2. Determine relevant passages based on unigram and bigram features. **Output** : relevant passages	Matching words using TF-IDF, thesaurus match using TF-IDF, mis-match words using TF-IDF, dispersion, cluster words. Next, the sentences ranked, and the top 5 sentences then selected as the answer.
[74]	They used IBM family of translation models [102]. **Output** : rank of candidate answers	They used Harmonic mean. The documents with the highest rank selected as the best answer.
[85]	Determine Named Entity Recognition (NER) using SVM, Baseline, Decision Tree (C4.5 & C5), and Maximum Entropy. **Output** : NER	Answer extraction based on NER.
[71]	They used SMART system [103] to retrieve paragraphs relevant to the target question. **Output** : paragraphs.	1. Identification of Relevant Sentences using semantic class for Named Entity. 2. Sort the sentences using Quickshort algorithm. 3. Compute a relative comparison between any two sentences using simple neural network. 4. The top 5 ranked sentences selected as the answers
[35]	They build a Semantic Interpreter using machine learning as in [104] that maps fragments of text into a weighted vector. Cosine similarity to selected the top scoring verses. **Output** : verses and their Tafseer.	1. Determine Named entities and question type. Next, measure the maximum count of named entity types. 2. Obtain the minimum distance between matched terms in the passage
[89]	Vector Space Model. **Output** : documents.	The answer selected from the highest score document.

Based on Table 9, most researchers used TF-IDF algorithm for terms weighting and Vector Space Model for measuring similarity between texts. Furthermore, to determine relevant passages in the document, they used N-gram technique. Then, to determine the answer, most researchers selected from the highest score documents or top *n* ranked documents.

Furthermore, Table 10 describes document processing for each study used semantic approach and the answer extraction techniques.

Table 10: Analysis Review of Document Processing using Semantic Approach and Answer Extraction Techniques

Cit.	Document Processing	Answer Extraction
[63]	1. Keyword-matching technique to retrieve documents and passages. 2. Passages are eliminated if don't satisfy the semantic constraints specified in the question **Output** : relevant documents	1. Named Entity Recognition, Question Type. 2. Answer ranking based on distance between keywords. 3. The best answer is selected based on the highest relevance scores
[81]	String similarity matching, generic lexical resources such as WordNet, and a domain-dependent lexicon.	AquaLog provides three mechanisms to generate an answer, i.e. And/or linking, Conditional link to a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	Output : Ontology compatible triples (Onto-Triple)	term, Conditional link to a triple.
[1]	SPARQL **Output** : answer candidates	1. Semantic search using Apache Lucene. 2. Word matching using Apache Jena Fuseki.
[64]	Execute TMR (Text Meaning Representation). **Output** : queries and data without event from fact repository.	TMR executed with COME event script to receive additional facts about the event from fact repository.
[72]	Analyze the question type, analyze keywords (noun and verbs), obtain the main structure, and retrieve similar patterns. **Output** : answer pairs related to the pattern	1. **Answer evaluation stage**. Calculate the matched parts, the weight of different parts of question, and the answer score 2. Measure the answer quality using bisecting K-means algorithm
[87]	Entailment Engine used *Lexical Inferences* and *Semantic ontology-based inferences*. To compute Lexical measures, they used Smith-Waterman, Consecutive subsequence matching, Jaro distance, Euclidean distance, and Jaccard similarity coefficien. **Output** : SPARQL.	Execute SPARQL
[77]	1. Keywords synonyms retrieval. 2. Determine ontology classes and their properties. 3. Build the appropriate SPARQL query. **Output** : DBpedia SPARQL query.	Execute SPARQL
[28]	Determine semantic associations between multiple ontology classes or properties using Semantic association discovery based on the Lowest Common Ancestor (LCA) and path finding. **Output** : RDF triples	1. Translates RDF triples into SPARQL using Cuebee [105]. 2. Execute SPARQL
[36]	1. Query relaxation approach. It is used to tackle annotation errors. 2. Semantic search approach by executed SPARQL query. **Output** : answer candidates	The answers are ranked according to two criteria: 1. Answers are ranked according to the queries rank 2. Their second ranking criteria for factual questions takes account of the number of justifications

Based on Table 10, most researchers determined ontology classes and their properties for the query, and determined RDF triples to build the appropriate SPARQL query and on document processing stage. Next, they executed SPARQL query to retrieve the answer from the knowledge repository.

Furthermore, Table 11 describes document processing for each study that used rule based approach and the answer extraction techniques.

Table 11: Analysis Review of Document Processing using Rule Based Approach and Answer Extraction Techniques

Cit.	Document Processing	Answer Extraction
[61]	Semantic rules using Knuth-style semantic interpreter. **Output** : LISP S-expression containing an array of functions.	Utilize the top level setx function from LISP S-expression to fetch a set of objects.
[68]	Case Frame Analysis and Domain Dependent Translation **Output** : Frame change descriptions and Current frame instances	Find the appropriate template and generates the English by filling in the template form
[70]	Tokenizer, entity identifiction and keyword detection, assign treebank POS tags to each token, morphological analysis, and add QLF to a semantic net. **Output** : The discourse model which derive from each sentence on QLF.	1. Sentence scoring 2. Entity scoring 3. The total score for every sentence
[79]	Semantic Matching technique. **Output** : One or more Minimal Logical Forms (MLFs).	Translated the MLFs into Prolog predicates and theorem prover to find the answers.
[69]	SQL Query Generation using breadth-first search to identify the complete set of queries. **Output** : SQL code	Execute SQL code
[66]	1. Question type identification based on pattern matching strings. 2. Checking for correlation between the pattern and the question's semantics. **Output** : Pattern matching string.	1. The exact answer identification in the pattern-matching string. 2. Calculate the total score for each candidate answer. 3. Selected the top-ranking candidate.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Based on Table 11, we could describe that every researcher has its own rule-based pattern. Their patterns conform to the case study and particular domain. The rule-based pattern in particular study is not necessarily used in another study. Next, to determine the answer, most researchers selected from the highest score documents or top n ranked documents.

Furthermore, Table 12 describes document processing for each study used hybrid approach and the answer extraction techniques.

Table 12: Analysis Review of Document Processing using Hybrid Approach and Answer Extraction Techniques

Cit.	Document Processing	Answer Extraction
[83]	1. Semantic search model (SSM) searching the Quranic ontology dataset by using SPARQL. If no result is found, then execute KSM 2. Keyword search model (KSM). Based on words matching technique **Output** : verses	1. Scoring and Ranking Model (SRM) eliminating the redundant verses from SSM and KSM. 2. SRM ranks and scores the refined results based on the number of matching words in the results.
[21]	This retrieval stage generally use Boolean methods, term weighting, and vector method. **Output** : relevant text passages	System ranking based on the semantic relations which calculate similarity between the question and the candidate answers.
[51]	1. Passage retrieval using the Lucene IR engine. 2. Passages are split into sentences and processed with POS tagging, chunking, and NERC. **Output** : The sentences list is composed of all named entities and all phrases containing a noun.	Answer ranking stages: 1. Context scores using BLEU [106] and ROGUE [107]. 2. Language scores. 3. A ranked list of answers.
[76]	1. Entity identification based on string similarity. String similarity measure using trigram, Levenshtein, and substring similarities. 2. Predicates detection using BOA framework [108]. **Output** : Entity and patterns for the property.	1. Rank the possible SPARQL queries using a similarity score and a prominence score. 2. Execute SPARQL with the highest score.
[7]	Semantic relationship identification using inferring schema mapping (ISM). This mapping consist of N-gram technique, Jaccard measure, attribute-based inference, and query simplification techniques. **Output** : SQL command	Execute SQL code

Based on Table 12, we could conclude that a hybrid approach is a combination of linguistic, statistical, semantic, and rule-based pattern matching techniques to retrieve answers from the data source.

Furthermore, Table 13 describes document processing for each study that used Google, MSN, ASK, Altavista, or other search engine, therewith the answer extraction techniques.

Table 13: Analysis Review of Document Processing using Existing Search Engine and Answer Extraction Techniques

Cit.	Document Processing	Answer Extraction
[80]	Google search engine	Google search engine
[41]	1. Retrieve documents through Google and MSN search engine. 2. Filter the documents with collect the first N hits, tokenization, and Decomposed text into sentences. **Output** : relevant sentences from each document	N-gram and statistical translation for answer extraction.
[75]	MSN, ASK, Google, Altavista, and Gigablast search engine. **Output** : documents.	Semantic roles using patterns and adding WordNet semantic classes to the patterns in order to filter the potential answers.
[86]	Retrieve the top twenty documents through Google search engine. **Output** : documents.	1. Question type identification (factoid or nonfactoid). 2. Calculate the similarity between the question and the document passages to return the best passages in a ranked list.
[78]	They used the MediaWiki API to retrieve 25 top ranked candidate entities from Wikipedia. **Output** : Wikipedia pages.	1. Split the Wikipedia category names to single words and remove all stopwords. 2. Stem the remaining words using the Porter stemming algorithm. 3. Vector standardization. 4. Entity clustering using k-means clustering. 5. Chose the winning cluster by select the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		cluster with the most data points in it.

Based on Table 13, after answering candidates retrieved through the search engine, most researchers used semantic or statistical approach for answer extraction. Next, to determine the best answer, most researchers selected from the highest score documents.

4. OPEN RESEARCH ISSUES

There are open research issues that can be highlighted for question analysis, document processing, and answer extraction techniques on Question Answering Systems:

A. Languages

Natural Language Processing (NLP) technique is applied at the pre-processing stage on question analysis processing. NLP is used to parse the text and to perform morphology analysis, such as sentences splitter, tokenizer, syntactic information provision or Part of Speech (POS) tagger, and to deduct a noun phrase (NP chunker). Every language has different written form, grammar, vocabulary, and syntax [109–111]. According to this condition, NLP technique for particular language has a method to perform morphological analysis which is different from other languages.

B. Question Classification

Supervised, semi supervised and unsupervised algorithm could be used to question classification. Study by [112] used Sequential Minimal Optimization, Naive Bayes, k-Nearest Neighbor, C4.5, and Random Forest algorithm for question classification. Study by [94] used SVM algorithm. Meanwhile, research by [92], [113] used Sparse Network of Winnows (SNoW) algorithm. Different approaches done by [114] used Semi-Bagging and Semi-AdaBoost, [93] used Question property kernel, and [115] used semantic approach for question classification.

According to the research results [86], SVM has a classification accuracy which is better than SNoW algorithm. They tested both algorithms with 3204 question training, 12 test set, 11 question class, factoid and nonfactoid question types. Research by [116] used data greater than [86]. They tested SVM algorithm, Naive Bayes, and Maximum Entropy with 3.1 million question training, 800 thousand test set, 54 question class, and simple question types. Their research results showed that SVM algorithm has a classification accuracy which is better than other algorithms. Some researchers tried to propose new or another algorithms to increase the question classification precision better than SVM. By using dataset from TREC 1999 – 2003 consisting of 5500 question training and 500 test set, research by [117] proposed Profile Hidden Markov Models (PHMMs) algorithm. They used 6 question classes. Their research results showed PHMMs accuracy was better than SVM. PHMMs precision was 92.2%, while SVM was between 90% and 91.8%. With similar dataset, study by [115] proposed semantic approach for question classification. They used 56 question class. Their research results showed that this approach has precision between 86.43% – 93%. Meanwhile, study by [95] proposed LibSVM algorithm. They used 56 question class. Their research results showed that this algorithm had 95% precision for coarse class and 90.8% for fine class.

Supervised learning, like SVM, usually requires a large training corpus to learn a classifier that performs well [118]. Shortcomings from supervised learning, if the dataset sized is small, then the accuracy of the classifier may decline [28][94][119]. Research results [119] showed that SVM classifier accuracy is weak in small dataset. Challenge in question classification is what technique could be used in small data set with a large question class (class labelling) for high classification accuracy. A large data set may permit more class labeling. However, a smaller data set necessitates fewer labels [28].

C. Term Weighting Using TF-IDF Algorithm

Based on the literature reviews, TF-IDF algorithm is used in several modules on question answering systems. In question analysis module, TF-IDF algorithm is used to calculate term weighting. Then, the calculation results from TF-IDF are used on question classification stage. Furthermore, the calculation result of TF-IDF is used in similarity measure on document processing module. In answer extraction module, TF-IDF is used to score and rank the answer candidates.

In classification stage, term weighting is the basis issues that affected the accuracy of classification results [120–124]. According to [120], [125-126], TF-IDF isn't effective algorithm for text classification, due to TF-IDF ignores the class labeling on training document.

5. CONCLUSION

In this study, the fundamental concepts and techniques related to question analysis, document processing, and answer extraction on question

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

answering systems have been discussed. The paper starts with introduction to question answering systems and provides past and present works found in the literature. Many research opportunities are still available along this line and further investigations for morphological analysis in a different language, question classification, and term weighting algorithm for question classification.

6. ACKNOWLEDGEMENT

The authors would like to thank the financial support from STMIK AMIKOM Purwokerto and Universiti Teknikal Malaysia Melaka for their assistance in this research.

REFERENCES:

[1] A. Sayed and A. Al Muqrishi, "An efficient and scalable Arabic semantic search engine based on a domain specific ontology and question answering," *Int. J. Web Inf. Syst.*, vol. 12, no. 2, pp. 242–262, 2012.

[2] S. Saqaeeyan, M. Shakibafakhr, and M. Roshanzadeh, "Proposed an Optimal Search Algorithm to Find the Best Answer in a Question Answering Systems," *Analele Univ. "Eftimie Murgu,"* vol. 21, no. 1, pp. 311–320, 2014.

[3] A. C. Mendes, L. Coheur, J. Silva, and H. Rodrigues, "JUST.ASK — A MULTI-PRONGED APPROACH TO QUESTION ANSWERING," *Int. J. Artif. Intell. Tools*, vol. 22, no. 1, pp. 1–34, 2013.

[4] O. Kolomiyets and M. F. Moens, "A survey on question answering technology from an information retrieval perspective," *Inf. Sci. (Ny).*, vol. 181, no. 24, pp. 5412–5434, 2011.

[5] D. Ferrés and H. Rodríguez, "Experiments Adapting an Open-Domain Question Answering System to the Geographical Domain Using Scope-Based Resources," in *Proceedings of the Workshop on Multilingual Question Answering*, 2006, pp. 69–76.

[6] R. Devi and M. Dua, "Performance Evaluation of Different Similarity Functions and Classification Methods Using Web Based Hindi Language Question Answering System," in *Procedia Computer Science*, 2016, vol. 92, pp. 520–525.

[7] A. Abdi, N. Idris, and Z. Ahmad, "QAPD: an ontology-based question answering system in the physics domain," *Soft Comput.*, pp. 1–18, 2016.

[8] M. Pavlić, Z. Dovedan Han, and A. Jakupović, "Question answering with a conceptual framework for knowledge-based system development 'node of Knowledge,'" *Expert Syst. Appl.*, vol. 42, no. 12, pp. 5264–5286, 2015.

[9] J. Peral, A. Ferrández, E. De Gregorio, J. Trujillo, A. Maté, and L. J. Ferrández, "Enrichment of the phenotypic and genotypic Data Warehouse analysis using Question Answering systems to facilitate the decision making process in cereal breeding programs," *Ecol. Inform.*, vol. 26, no. 2, pp. 203–216, 2015.

[10] J. Bian, Y. Liu, E. Agichtein, and H. Zha, "Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 467–476.

[11] D. Hristovski, D. Dinevski, A. Kastrin, and T. C. Rindflesch, "Biomedical question answering using semantic relations," *BMC Bioinformatics*, vol. 16, p. 6, 2015.

[12] Jovita, Linda, A. Hartawan, and D. Suhartono, "Using Vector Space Model in Question Answering System," in *Procedia Computer Science*, 2015, vol. 59, pp. 305–311.

[13] R. F. Simmons, "Natural Language Question-answering Systems: 1969," *Commun. ACM*, vol. 13, no. 1, pp. 15–30, 1970.

[14] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[15] A. B. Ellis and D. V Tiedeman, "Can A Machine Counsel?," in *CEEB-SSRC Conference on Computer-Based Instruction, Learning, Testing, and Guidance*, 1968, p. 41.

[16] K. M. Colby and H. Enea, "Inductive inference by intelligent machines," California, 1968.

[17] R. C. Schank and L. G. Tesler, "A CONCEPTUAL PARSER FOR NATURAL LANGUAGE," in *Proceedings of the First International Joint Conference*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

on Artificial Intelligence, 1969, pp. 569–578.

[18] I. Gurevych, D. Bernhard, K. Ignatova, and C. Toprak, "Educational Question Answering based on Social Media Content," in *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, 2009, vol. 200, pp. 133–140.

[19] Z. Liu and B. J. Jansen, "Understanding and Predicting Question Subjectivity in Social Question and Answering," *IEEE Trans. Comput. Soc. Syst.*, vol. 3, no. 1, pp. 32–41, 2016.

[20] W. Chen, "Developing a Framework for Geographic Question Answering Systems Using GIS, Natural Language Processing, Machine Learning, and Ontologies," Ohio State University, 2014.

[21] A. Mishra, N. Mishra, and A. Agrawal, "Context-aware restricted geographical domain question answering system," in *Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks*, 2010, pp. 548–553.

[22] G. Cai, "Contextualization of geospatial database semantics for human-GIS interaction," *Geoinformatica*, vol. 11, no. 2, pp. 217–237, 2007.

[23] J. L. Leidner, G. Sinclair, and B. Webber, "Grounding spatial named entities for information extraction and question answering," in *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, 2003, vol. 1, pp. 31–38.

[24] W. A. Woods, "Progress in natural language understanding: an application to lunar geology," in *Proceedings of the National Computer Conference and Exposition on AFIPS '73*, 1973, pp. 441–450.

[25] M. Serhatli and F. Alpaslan, "An Ontology Based Question Answering System on Software Test Document Domain," *Int. J. Comput. Electr. Autom. Control Inf. Eng.*, vol. 3, no. 6, pp. 1630–1634, 2009.

[26] J. Cheng, B. Kumar, and K. H. Law, "A question answering system for project management applications," *Adv. Eng. Informatics*, vol. 16, no. 4, pp. 277–289, 2002.

[27] D. L. Waltz, "An English language question answering system for a large relational database," *Commun. ACM*, vol. 21, no. 7, pp. 526–539, 1978.

[28] A. H. Asiaee, P. Doshi, T. Minning, and R. L. Tarleton, "A framework for ontology-based question answering with application to parasite immunology," *J. Biomed. Semantics*, vol. 6, p. 31, 2015.

[29] W. Wong, J. Thangarajah, and L. Padgham, "Contextual Question Answering for the Health Domain," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 11, pp. 2313–2327, 2012.

[30] Y. Cao *et al.*, "AskHERMES: An online question answering system for complex clinical questions," *J. Biomed. Inform.*, vol. 44, no. 2, pp. 277–288, 2011.

[31] R. T. K. Lin, J. L. Chiu, H. Dai, R. T. Tsai, M.-Y. Day, and W.-L. Hsu, "A Supervised Learning Approach to Biological Question Answering," *Integr. Comput. Aided. Eng.*, vol. 16, no. 3, pp. 271–281, 2009.

[32] A. G. Tapeh and M. Rahgozar, "A knowledge-based question answering system for B2C eCommerce," *Knowledge-Based Syst.*, vol. 21, no. 8, pp. 946–950, 2008.

[33] S. K. Hamed and M. J. A. Aziz, "A question answering system on Holy Quran translation based on question expansion technique and Neural Network classification," *J. Comput. Sci.*, vol. 12, no. 3, pp. 169–177, 2016.

[34] A. Hakkoum and S. Raghay, "Semantic Q&A System on the Qur'an," *Arab. J. Sci. Eng.*, vol. 41, no. 12, pp. 5205–5214, 2016.

[35] H. Abdelnasser *et al.*, "Al-Bayan : An Arabic Question Answering System for the Holy Quran," in *9th International Workshop on Semantic Evaluation*, 2014, pp. 57–64.

[36] A. Ben Abacha and P. Zweigenbaum, "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies," *Inf. Process. Manag.*, vol. 51, no. 5, pp. 570–594, 2015.

[37] R. Barskar, G. F. Ahmed, and N. Barskar, "An approach for extracting exact answers to Question Answering (QA) system for english sentences," in *Procedia Engineering*, 2012, vol. 30, pp. 1187–1194.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[38] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question Answering Systems: Survey and Trends," in *Procedia Computer Science*, 2015, vol. 73, pp. 366–375.

[39] S. K. Dwivedi and V. Singh, "Research and Reviews in Question Answering System," in *Procedia Technology*, 2013, vol. 10, pp. 417–424.

[40] K. Komiya, Y. Abe, H. Morita, and Y. Kotani, "Question answering system using Q & A site corpus Query expansion and answer candidate evaluation.," *Springerplus*, vol. 2, p. 396, 2013.

[41] R. Soricut and E. Brill, "Automatic question answering using the web: Beyond the Factoid," *Inf. Retr. Boston.*, vol. 9, no. 2, pp. 191–206, 2006.

[42] A. Mishra and S. K. Jain, "A survey on question answering systems with classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 28, no. 3, pp. 345–361, 2015.

[43] D. Ferres *et al.*, "The TALP-QA system for Spanish at CLEF 2004: Structural and hierarchical relaxing of semantic constraints," *Multiling. Inf. Access Text, Speech Images*, vol. 3491, pp. 557–568, 2005.

[44] S. K. Ray and K. Shaalan, "A Review and Future Perspectives of Arabic Question Answering Systems," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3169–3190, 2016.

[45] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton, "Quantitative evaluation of passage retrieval algorithms for question answering," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 2003, pp. 41–47.

[46] D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal, "Probabilistic question answering on the Web," *J. Am. Soc. Inf. Sci. Technol.*, vol. 56, no. 6, pp. 571–583, 2005.

[47] M. Vargas-Vera and M. D. Lytras, "AQUA: Hybrid architecture for question answering services," *IET Softw.*, vol. 4, no. 6, pp. 418–433, 2010.

[48] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," *Nat. Lang. Eng.*, vol. 7, no. 4, pp. 275–300, 2001.

[49] D. Tufis, "Natural Language Question Answering in Open Domains," *Comput. Sci. J. Mold.*, vol. 19, no. 2, pp. 146–164, 2011.

[50] D. Carmel, A. Shtok, and O. Kurland, "Position-based contextualization for passage retrieval," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 2013, pp. 1241–1244.

[51] C. España-Bonet and P. R. Comas, "Full Machine Translation for Factoid Question Answering," in *EACL Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, 2012, pp. 20–29.

[52] J. Lin, "An exploration of the principles underlying redundancy-based factoid question answering," *ACM Trans. Inf. Syst.*, vol. 25, no. 2, pp. 1–55, 2007.

[53] P. R. Comas, J. Turmo, and L. Marquez, "Sibyl, a Factoid Question-Answering System for Spoken Documents," *Acm Trans. Inf. Syst.*, vol. 30, no. 3, pp. 1–40, 2012.

[54] I. B. M. Watson *et al.*, "Fact-based question decomposition in DeepQA," *IBM J. Res. Dev.*, vol. 56, no. 3.4, p. 13:1-13:11, 2012.

[55] M. M. Hoque and P. Quaresma, "SEMONTOQA: A Semantic Understanding-Based Ontological Framework for Factoid Question Answering," in *Proceedings of the Forum for Information Retrieval Evaluation on - FIRE '14*, 2014, pp. 10–20.

[56] J. Fukumoto, "Question Answering System for Non-factoid Type Questions and Automatic Evaluation based on BE Method," in *Proceedings of NTCIR-6 Workshop Meeting*, 2007, pp. 441–447.

[57] T. Mori, M. Sato, M. Ishioroshi, Y. Nishikawa, S. Nakano, and K. Kimura, "A Monolithic Approach and a Type-by-Type Approach for Non-Factoid Question-answering," in *Proceedings of NTCIR-6 Workshop Meeting*, 2007, pp. 469–476.

[58] T. Mori, M. Sato, and M. Ishioroshi, "Answering any class of Japanese non-factoid question by using the Web and example Q & A pairs from a social Q & A website," in *Proceedings - 2008 IEEE/WIC/ACM International Conference*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

on Web Intelligence, 2008, pp. 59–65.

[59] H. Shima and T. Mitamura, "JAVELIN III: Answering Non-Factoid Questions in Japanese," in *Proceedings of NTICIR-6 Workshop*, 2007, pp. 464–468.

[60] P. Achananuparp, C. C. Yang, and X. Chen, "Using Negative Voting to Diversify Answers in Non-Factoid Question Answering," in *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*, 2009, pp. 1681–1684.

[61] W. J. Plath, "REQUEST: A Natural Language Question-Answering System," *IBM J. Res. Dev.*, vol. 20, no. 4, pp. 326–335, 1976.

[62] M. Mishra, V. K. Mishra, and H. R. Sharma, "Performance measurement for the quality of question answering approaches in natural language," in *2012 2nd National Conference on Computational Intelligence and Signal Processing (CISP)*, 2012, pp. 95–98.

[63] D. Moldovan *et al.*, "LCC Tools for Question Answering," in *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, 2002, pp. 388–397.

[64] S. Beale, B. Lavoie, M. McShane, S. Nirenburg, and T. Korelsky, "Question answering using ontological semantics," in *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*, 2004, pp. 41–48.

[65] M. M. Soubboutin, "Patterns of Potential Answer Expressions as Clues to the Right Answers," in *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 2001, pp. 293–302.

[66] M. M. Soubbotin and S. M. Soubbotin, "Use of patterns for detection of answer strings: A systematic approach," in *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, 2002, vol. 52, p. 90.

[67] S. I. A. Saany, A. Mamat, A. Mustapha, L. S. Affendey, and M. N. A. Rahman, "Semantics Question Analysis Model for Question Answering System," *Appl. Math. Sci.*, vol. 9, no. 130, pp. 6491–6505, 2015.

[68] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd, "GUS, a frame-driven dialog system," *Artif. Intell.*, vol. 8, no. 2, pp. 155–173, 1977.

[69] P. Pruski, S. Lohar, W. Goss, A. Rasin, and J. Cleland-Huang, "TiQi: Answering unstructured natural language trace queries," *Requir. Eng.*, vol. 20, no. 3, pp. 215–232, 2015.

[70] S. Scott and R. Gaizauskas, "University of Sheffield TREC-9 Q&A System," in *Proceedings of the Ninth Text REtrieval Conference (TREC 2000)*, 2000, pp. 635–644.

[71] A. Moschitti, "Answer Filtering via Text Categorization in Question Answering Systems," in *Proc. of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2003)*, 2003, pp. 241–248.

[72] T. Hao, D. Hu, L. Wenyin, and Q. Zeng, "Semantic patterns for user-interactive question answering," *Concurr. Comput. Pract. Exp.*, vol. 20, no. 7, pp. 751–902, 2008.

[73] A. R. Diekema, J. Chen, N. Mccracken, N. E. Ozgencil, and M. D. Taffet, "Question Answering: CNLP at the TREC-2002 Question Answering Track," in *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, 2002.

[74] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal, "Bridging the lexical chasm: statistical approaches to answer-finding," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000, pp. 192–199.

[75] P. Moreda, H. Llorens, E. Saquete, and M. Palomar, "Combining semantic information in question answering systems," *Inf. Process. Manag.*, vol. 47, no. 6, pp. 870–885, 2011.

[76] C. Unger, L. Bühmann, J. Lehmann, A.-C. N. Ngomo, D. Gerber, and P. Cimiano, "Template-based question answering over RDF data," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 639–648.

[77] A. Tahri and O. Tibermacine, "Dbpedia Based Factoid Question Answering System," *Int. J. Web Semant. Technol.*, vol. 4, no. 3, pp. 23–38, 2013.

[78] G. Stratogiannis, G. Siolas, and A. Stafylopatis, "Semantic Question Answering Using Wikipedia Categories Clustering," *Int. J. Artif. Intell. Tools*, vol. 23, no. 4, pp. 1–20, 2014.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[79] D. Mollá, R. Schwitter, F. Rinaldi, J. Dowdall, and M. Hess, "NLP for answer extraction in technical domains," in *Proceedings of The Workshop on Natural Language Processing for Question Answering*, 2003.

[80] L. Wu, X. Huang, Y. Zhou, Y. Du, and L. You, "FDUQA on TREC2003 QA task," in *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, 2003, pp. 246–253.

[81] V. Lopez, V. Uren, E. Motta, and M. Pasin, "AquaLog: An ontology-driven question answering system for organizational semantic intranets," *Web Semant.*, vol. 5, no. 2, pp. 72–105, 2007.

[82] S. M. Archana, N. Vahab, R. Thankappan, and C. Raseek, "A Rule Based Question Answering System in Malayalam Corpus Using Vibhakthi and POS Tag Analysis," in *Procedia Technology*, 2016, vol. 24, pp. 1534–1541.

[83] M. Alqahtani and E. Atwell, "Arabic Quranic Search Tool Based on Ontology," in *21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016*, 2016, vol. 9612, pp. 478–485.

[84] A. Ittycheriah, M. Franz, W.-J. Zhu, and A. Ratnaparkhi, "IBM's Statistical Question Answering System," in *Proceedings of the Ninth Text REtrieval Conference (TREC 2000)*, 2000, pp. 229–234.

[85] J. Suzuki, Y. Sasaki, and E. Maeda, "SVM answer selection for open domain question answering," in *Proceedings of the 19th international conference on Computational linguistics*, 2002, pp. 1–7.

[86] S. Quarteroni and S. Manandhar, "Designing an interactive open-domain question answering system," *Nat. Lang. Eng.*, vol. 15, no. 1, pp. 73–95, 2009.

[87] Ó. Ferrández, R. Izquierdo, S. Ferrández, and J. L. Vicedo, "Addressing ontology-based question answering with collections of user queries," *Inf. Process. Manag.*, vol. 45, no. 2, pp. 175–188, 2009.

[88] R. H. Gusmita, Y. Durachman, S. Harun, A. F. Firmansyah, H. T. Sukmana, and A. Suhaimi, "A rule-based question answering system on relevant documents of Indonesian Quran Translation," in *2014 International Conference on Cyber and IT Service Management, CITSM 2014*, 2014, pp. 104–107.

[89] E. Al-Shawakfa, "A RULE-BASED APPROACH TO UNDERSTAND QUESTIONS IN ARABIC QUESTION ANSWERING," *Jordanian J. Comput. Inf. Technol.*, vol. 2, no. 3, pp. 210–231, 2016.

[90] D. D. K. Sleator and D. Temperley, "Parsing English with a Link Grammar," in *Third International Workshop on Parsing Technologies*, 1995, vol. 64, no. October, pp. 1–91.

[91] H. Cunningham, "GATE, a general architecture for text engineering," *Comput. Hum.*, vol. 36, no. 2, pp. 223–254, 2002.

[92] X. Li and R. Dan, "Learning Question Classifiers," in *COLING '02 Proceedings of the 19th international conference on Computational linguistics*, 2002, pp. 1–7.

[93] L. Liu, Z. Yu, J. Guo, C. Mao, and X. Hong, "Chinese Question Classification Based on Question Property Kernel," *Int. J. Mach. Learn. Cybern.*, vol. 5, no. 5, pp. 713–720, 2013.

[94] D. Metzler and W. B. Croft, "Analysis of statistical question classification for fact-based questions," *Inf. Retr. Boston.*, vol. 8, no. 3, pp. 481–504, 2005.

[95] J. Silva, L. Coheur, A. C. Mendes, and A. Wichert, "From symbolic to sub-symbolic information in question classification," *Artif. Intell. Rev.*, vol. 35, no. 2, pp. 137–154, 2011.

[96] P. R. Cohen, "The role of natural language in a multimodal interface," in *Proceedings of the ACM symposium on User interface software and technology (UIST '92)*, 1992, pp. 143–149.

[97] A. M. Pundge, S. . Khillare, and C. N. Mahender, "Question Answering System, Approaches and Techniques: A Review," *Int. J. Comput. Appl.*, vol. 141, no. 3, pp. 34–39, 2016.

[98] D. L. McGuinness, "Question answering on the semantic Web," *IEEE Intell. Syst.*, vol. 19, no. 1, pp. 82–85, 2004.

[99] D. L. McGuinness, "Ontological Issues for Knowledge-Enhanced Search," in *Proceedings of The First International Conference (FOIS'98)*, 1998, pp. 302–316.

[100] L. A. Pizzato and D. Mollá, "Question Prediction Language Model," in *Proceedings of the Australasian Language Technology Workshop 2007*, 2007, pp. 92–

99.

[101] S. I. A. Saany, A. Mamat, A. Mustapha, and L. S. Affendey, "A Strategy for Question Interpretation in Question Answering System," *Int. J. Comput. Sci. Telecommun.*, vol. 4, no. 5, pp. 38–43, 2013.

[102] P. Brown *et al.*, "A statistical approach to machine translation," *Comput. Linguist.*, vol. 16, no. 2, pp. 79–85, 1990.

[103] J. J. Rocchio, "Relevance Feedback in Information Retrieval," Cambridge, 1971.

[104] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," in *Proceedings of the 20th international joint conference on Artifical intelligence*, 2007, pp. 1606–1611.

[105] P. N. Mendes, B. McKnight, A. P. Sheth, and J. C. Kissinger, "TcruziKB: Enabling complex queries for genomic data exploration," in *Proceedings - IEEE International Conference on Semantic Computing 2008, ICSC 2008*, 2008, pp. 432–439.

[106] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, no. July, pp. 311–318.

[107] C. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004, pp. 1–8.

[108] D. Gerber and A. N. Ngomo, "Bootstrapping the Linked Data Web," in *1st Workshop on Web Scale Knowledge Extraction@ ISWC*, 2011.

[109] B. Andi-pallawa, "A Comparative Analysis between English and Indonesian Phonological Systems," *Int. J. English Lang. Educ.*, vol. 1, no. 3, pp. 103–129, 2013.

[110] B. R. Chiswick and P. W. Miller, "Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages," *J. Multiling. Multicult. Dev.*, vol. 26, no. August, pp. 1–18, 2005.

[111] A. U. Rahayu, "Differences on Language Structure between English and Indonesian," *Int. J. Lang. Lit. Linguist.*, vol. 1, no. 4, pp. 257–260, 2015.

[112] S. Basuki and A. Purwarianti, "Statistical-based approach for Indonesian complex factoid question decomposition," *Int. J. Electr. Eng. Informatics*, vol. 8, no. 2, pp. 356–373, 2016.

[113] Y. Chali, "Question Answering Using Question Classification and Document Tagging," *Appl. Artif. Intell.*, vol. 23, no. 6, pp. 500–521, 2009.

[114] Y. Li, L. Su, J. Chen, and L. Yuan, "Semi-supervised learning for question classification in CQA," *Nat. Comput.*, pp. 1–11, 2016.

[115] S. K. Ray, S. Singh, and B. P. Joshi, "A semantic approach for question classification using WordNet and Wikipedia," *Pattern Recognit. Lett.*, vol. 31, no. 13, pp. 1935–1943, 2010.

[116] B. Qu, G. Cong, C. Li, A. Sun, and H. Chen, "An evaluation of classification models for question topic categorization," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 5, pp. 889–903, 2012.

[117] Y. Pan, Y. Tang, Y. M. Luo, L. X. Lin, and G. B. Wu, "Question Classification Using Profile Hidden Markov Models," *Int. J. Artif. Intell. Tools*, vol. 19, no. 1, pp. 121–131, 2010.

[118] K. Zhang and J. Zhao, "A Chinese question-answering system with question classification and answer clustering," in *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010*, 2010, vol. 6, pp. 2692–2696.

[119] D. Tomás and J. L. Vicedo, "Minimally supervised question classification on fine-grained taxonomies," *Knowl. Inf. Syst.*, vol. 36, no. 2, pp. 303–334, 2013.

[120] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Syst. Appl.*, vol. 66, pp. 245–260, 2016.

[121] T. Sabbah, A. Selamat, M. H. Selamat, R. Ibrahim, and H. Fujita, "Hybridized term-weighting method for Dark Web classification," *Neurocomputing*, vol. 173, pp. 1908–1926, 2016.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[122] M. A. Fattah, "New term weighting schemes with combination of multiple classifiers for sentiment analysis," *Neurocomputing*, vol. 167, pp. 434–442, 2015.

[123] H. J. Escalante *et al.*, "Term-weighting learning via genetic programming for text classification," *Knowledge-Based Syst.*, vol. 83, no. 1, pp. 176–189, 2015.

[124] M. Kumari, A. Jain, and A. Bhatia, "Synonyms Based Term Weighting Scheme: An Extension to TF.IDF," in *Procedia Computer Science*, 2016, vol. 89, pp. 555–561.

[125] T. Peng, L. Liu, and W. Zuo, "PU text classification enhanced by term frequency–inverse document frequency-improved weighting," *Concurr. Comput. Pract. Exp.*, vol. 26, no. 3, pp. 728–741, 2013.

[126] Z. H. Deng, K. H. Luo, and H. L. Yu, "A study of supervised term weighting scheme for sentiment analysis," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3506–3513, 2014.