IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 8, August 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908029 |

Answer Extraction to QA System for Agriculture Domain (Fruit) using NLP

Mr. Santosha S. Peerappagol, **Dr. Vijay S. Rajpurohit**

PG Student, Dept. of CSE, KLS's Gogte Institute of Technology, Belagavi, India

HOD, Dept. of CSE, KLS's Gogte Institute of Technology, Belagavi, India

ABSTRACT: Question Answering (QA) research may be the most important and difficult task in Natural Language Processing. QA aims to extract an explicit answer from a relevant text document. Agriculture is the main source of a country's economic growth. Now a days farmers aren't abundant awake to recent technologies and practices being employed within the agriculture field. Extraction of meaningful answers by machine learning techniques is a problem, that has been studied by many machine learning experts as well as advanced machine learning techniques are introduced. These techniques are applied to extract the exact answer. Question Answering System for agriculture, where the farmer can query the system and the system understands the query and responds to a given query. This paper has reviewed extracting the exact answer for a given question by focusing on machine learning techniques. We have created suggestions and provided a comparative analysis.

KEYWORDS: Question Answering System, Natural Language Processing, Machine learning, Exact answer.

I. INTRODUCTION

From the last decade, internet users are increasing in more percentage because of availability of the information on the web. In QA different queries are provided by the user with aim of obtaining correct answers in Question Answering Systems. Question Answering provides the right solution to retrieve valid and correct answers to user queries asked in natural language rather than a question. Question answering may be a specialized space within the field of information extraction. Many question answering systems are having their application area. Question Answering System (QAS) has many applications based on the source of answers. Like extracting information from the document, language learning, online examination system, human and computer interaction, document management, classification of the document, and many more[1].

There has been significant research in the field of QA systems but there isn't any agriculture domain-specific QA system that returns actual answers by analyzing data to the questions posed by the farmers. Agriculture is the art and science of soil cultivation; it is the key to the development of the country. Today modern way of plant breeding, agrochemicals like pesticides, fertilizers and due to increased technological developments, our farmers are left behind, so they should take a step ahead. Making them aware of regarding latest techniques is necessary. Question Answering System is also called as human-computer interaction. The user queries the machine it should respond with accurate answers. Significant research is made on question answering systems in various domains such as medicine and travel.
Considering agriculture as a specific domain some systems are capable of exploring the web data, during this process, retrieving the exact answer is not possible. Since the website also leads to certain unrelated answers to the asked queries. To get the exact response we make use of Natural Language Processing, and Recurrent Neural Network (RNN) deep learning techniques. Hence, a study is to help farmers query regarding the diseases, crops, raw materials used, plants grown in a particular area, usage of pesticides and fertilizers, etc. Question answering system is built using Natural language processing and information extraction techniques. The QA system is divided into four modules namely. Question Processing Module, Document Processing Module, Paragraph Extraction Module, and Answer Extraction Module [1].

In Question processing module user query in natural language is processed using POS tagging, stemming and removal of keywords [5]. The document processing module retrieves the relevant set of documents from the Internet using any search engine. The paragraph extraction module performs the task of Paragraph Extraction and Sentence Extraction to find out the most probable answer to the question at hand. In answer extraction module retrieves the answer and tests the answer for the correctness and provides an exact answer to the user.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908029 |

II. LITERATURE SURVEY

Literature survey has been discussed in this section to highlight the work carried out till now in the QA system. It works mainly under three phases Question Identification, Knowledge Base searching, and providing exact answers.
AGRI-QAS [2] focuses on processing unstructured data and provides responses for FACTOID queries such as 'which', 'what', 'who', 'where'. Question Answering system for restricted domain [3] uses advanced NLP tools to work with information extraction technique. It presents research in a restricted domain that classifies the questions semantically into EAT's (Expected answer type), Headword of the questions, and question keywords. EAT's and headwords are found using different algorithms for definition type, descriptive type, and factoid type questions. ADANS[4] (An Agriculture Domain Question Answering System using Ontologies) represents a response to the question provided in natural language. NLP (Natural Language Processing) and semantic web technologies are used. SPARQL is formulated by the system, from the questions represented in natural language [4].

1. Question Identification

Question processing in AGRI-QAS [2] is defined by specific rules for pre-processing and post-processing. Adding hyphens between consecutive names, replacing two words with a similar word are some pre-processing rules. Post-processing rules are collaborated as the two synonyms and are interchangeable. Question Answering system for restricted domain processes questions in four steps. Firstly find out the question type using the 'Wh' word, find out the expected answer type, getting the keyword from the question, and find out the headword of the question [3].
In ADANS [4] the question processing is done in steps. It initiates with pre-processing of queries where tokenization, stop words removal, and POS tagging and stemming take place. After the completion of pre-processing, the next step is to form triples, using the Stanford dependency tree. The dependency tree provides a relationship between words in a sentence as subject, determiner, etc. To get the desired triple edges of the dependency tree are removed and merged [4].

2. Knowledge Base

AGRI-QAS [2] takes input as XML documents (news articles, blogs, etc) and grammar parser, POS tagging, named entity recognizer (NER) are implemented. Instead of tagging words as part of speech, indexing the documents according to domain-specific terms is carried out using a domain-specific named entity recognizer [2].
Question Answering system for restricted domain uses Alchemy Content Extractor to remove unimportant content from the webpages. Paragraph Extractor is used to extract only paragraphs which have the same keyword as the question keyword [3].
ADANS[4] uses an ontological Knowledge base. Ontology building is done using the tool Protégé. It is widely used for ontology construction. Firstly domain is decided, then different entities and their relationships are identified.

3. Answer Extraction

AGRI-QAS [2] forms a query by eliminating stop words from the question. LUCENE query uses the eliminated query to retrieve the answer by searching through the LUCENE index.
Question Answering system for restricted domain [3] uses Stanford's core NLP toolkit to find out the grammatical structure of the given question or sentence. Answer extraction is performed by comparing with keywords and headwords. The ADANS system uses SPARQL is a language used for querying data in RDF (Resource Description Framework) format [4]. SPARQL query is generated using the rule-based technique, stop words are removed from the triples. Then relation list from the ontology is extracted using EAT and subject.

III. RESEARCH GAPS

AGRI-QAS is designed to ensure flexibility. But the system does not support list-type questions[2]. The major problem of the QA system for a restricted domain is its performance dependence on the performance of the search engine and used NLP tools[3].
According to the survey, SPARQL query is used to extract answers [4]. The major drawback is, it can't hold dynamic data that gets updated frequently. SPARQL works well in a closed environment. It is time-consuming and difficult to handle negation statements like not, don't, etc.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908029 |

Title	Author	Technique Used	Year	Disadvantages
QAS [1]	Amit Kumar Punde, Khillare S. A, Namrata Mahender	Rule-based and NLP technique	2016	A rule-based approach is used for semi-structured and structured text documents. This approach includes predefined rules hence, learning is reduced.
AGRI-QAS [2]	Sharavati Gaikawad, Rohan Asodekar, Sunny Gadia, Vahida Attar	NLP and IR-based techniques	2015	AGRI-QAS is designed to ensure flexibility. But the system does not support list-type questions.
A Frame work for restricted domain QA system [3]	Payal Biswas, Aditi Sharan, Nidhi Malik	NLP and IE-based techniques	2014	The major problem of the QA system for a restricted domain is its performance dependence on the performance of the search engine and used NLP tools.
ADANS [4]	Manmita Devi, Mohit Dua	NLP and SPARQL techniques	2017	The major drawback is, SPARQL can't hold dynamic data that gets updated frequently. It is time-consuming and difficult to handle negation statements like not, don't, etc.
QA system based on Data Mining [5]	Wahid Ahmed, Babu Anto P	Text Mining technique and Supervised ML approach	2017	Lexical information requires a lot of training data to detect the relevant passages and the candidate answers.

- **Rule-based and NLP technique**

A rule-based approach is used for semi-structured and structured text documents. This approach includes predefined rules hence, learning is reduced [1]. There are no rules defined for every situation, so it is also time-consuming. In the NLP approach automatically rules are defined hence it is fast [5].

- **Supervised machine learning**

Various supervised machine learning algorithms are used as classifiers [1]. The decision tree is a simple algorithm used for classification. But, this technique doesn't produce efficient performance and accuracy. K-nearest neighbor has a high computational cost. When compared to SVM, Naïve Bayes has low efficiency. SVM performed better when compared to all supervised learning classification techniques. One disadvantage of SVM is most of the system memory is consumed and the complexity of the algorithm. Due to this, it can be suggested to use neural networks.

- **Statistical Approach**

This technique is called a "bag of words". It plays important role in web data and online platforms. The document contains various words that are identified by a group of keywords and based on the frequency weight is assigned to each word [1]. The drawback is every term is treated separately and it has failed to characterize linguistic properties for a group of phrases and terms [6].

IV. PROPOSED SYSTEM

Numerous advancements have been made in the fields of NLP and information retrieval during the past two decades. A number of tools and techniques have been proposed in order to improve the performance of the searching and retrieval process, including the development of a better similarity measure, the efficient design of an information retrieval

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908029 |

system, the availability of better resources and tools, such as POS taggers, NE recognizers, content extractors, and numerous others.

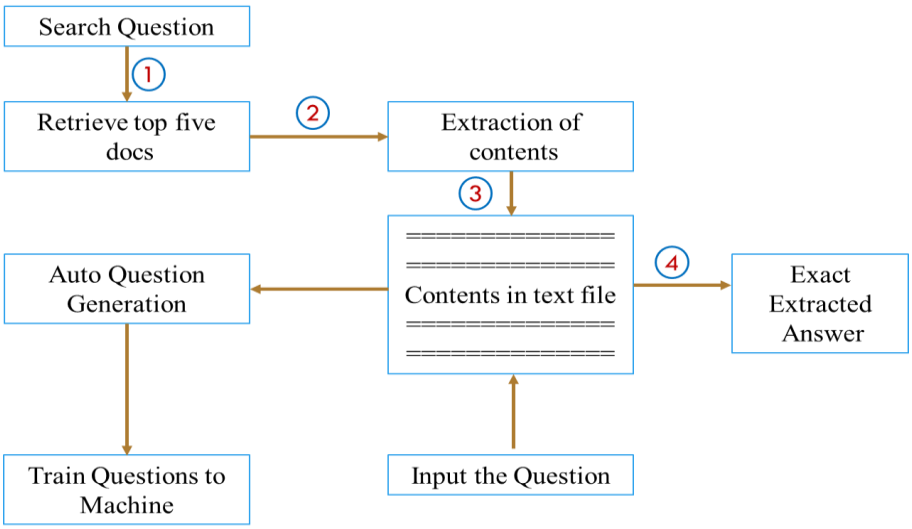

Fig 1: Proposed architecture of QA System

In light of these considerations, we have suggested a new design for the quality assurance system, which incorporates the newly created tools and methods, thus increasing the overall performance of the question answering system. The suggested design of the question answering system is shown in Fig 1.

1. **Question Processing Module**: In this, user inputs the query in the search engine, which lists the the various top ranked links for the query.
2. **Document Processing Module**: In this, get the top three links and extract the contents from each links and store it as separate files.
3. **Paragraph Extraction Module**: Now, merge the contents of each extracted files into a single file and then by using NLP generate the questions automatically and train the model.
4. **Answer Extraction Module**: By using RNN algorithm, exact answer are extracted. RNN encoder and RNN decoder are two major components of sequence to sequence algorithm. When the user provides input that is a query posed then required entities are extracted. RNN sequence to sequence algorithm is used which considers previous output to predict the answer.

The overall steps for QA System is shown in the flow diagram (see Fig 2).

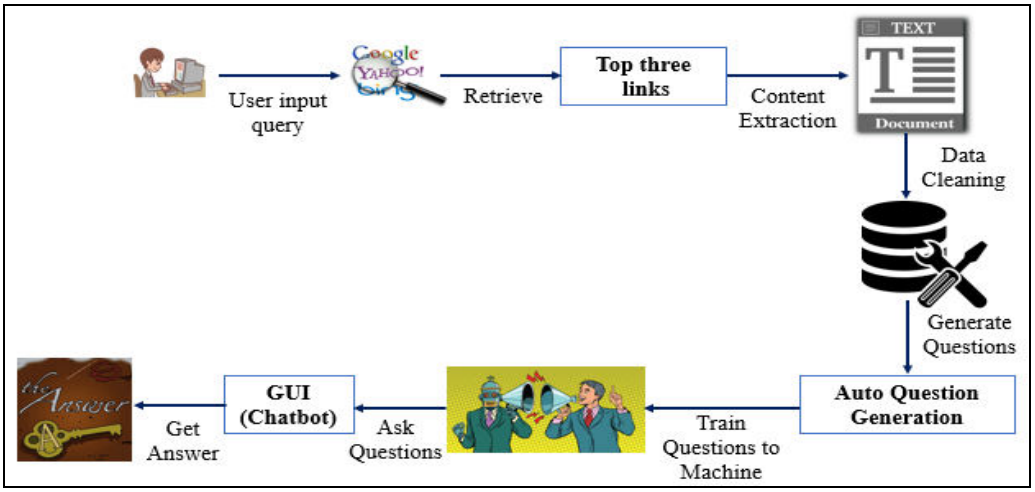

Fig 2: Flow diagram for QA System

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908029 |

V. RESULTS AND DISCUSSIONS

This section briefs about the results obtained in the phase of implementation of algorithm. The below figures show the step-by-step execution. Using the web pages the data is going to be extracted and the same is as shown in the Fig 3.

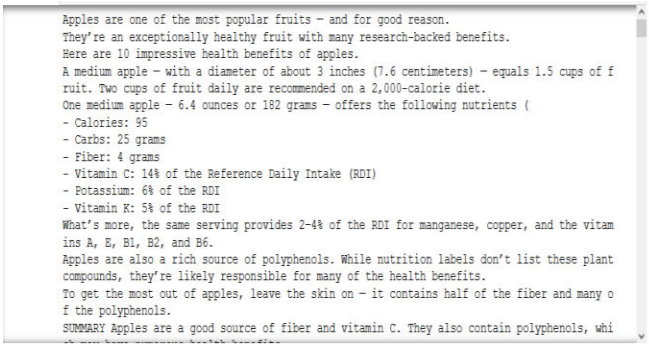

Fig 3: Online Data Extraction from top links

The extracted contents of all top links are merged into single extracted text file and the same is shown in below Fig 4.

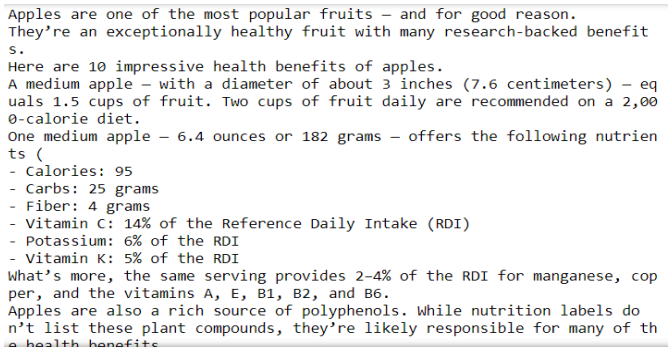

Fig 4: Data Extracted from top links are merged

The algorithm will generate the questions on its own by observing the dataset and the results are as shown in the Fig 5.

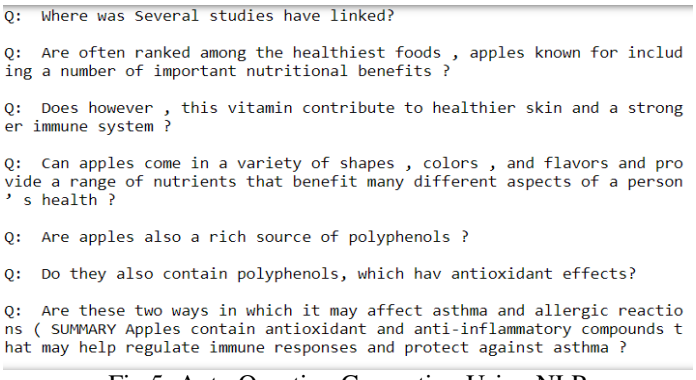

Fig 5: Auto Question Generation Using NLP

The model is going to train using the Recurrent Neural Network, the data is taken from the database and it is trained.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908029 |

```
uracy: 0.7017
Epoch 195/200
29/29 [==============================] - 0s 1ms/step - loss: 1.4347 - acc
uracy: 0.6557
Epoch 196/200
29/29 [==============================] - 0s 1ms/step - loss: 1.0655 - acc
uracy: 0.7785
Epoch 197/200
29/29 [==============================] - 0s 1ms/step - loss: 1.4120 - acc
uracy: 0.6985
Epoch 198/200
29/29 [==============================] - 0s 1ms/step - loss: 1.2859 - acc
uracy: 0.6433
Epoch 199/200
29/29 [==============================] - 0s 1ms/step - loss: 0.8567 - acc
uracy: 0.7995
Epoch 200/200
29/29 [==============================] - 0s 1ms/step - loss: 0.9959 - acc
uracy: 0.7378
model created
```

Fig 6: Training to RNN Model

The figure 7, 8 and 9 shows the process of extraction of questions, answers using the trained model.

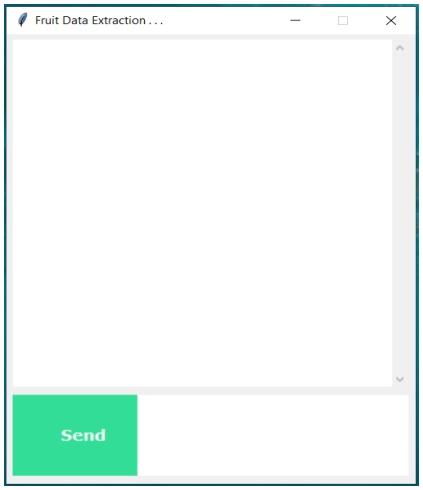

Fig 9: Extraction of QA

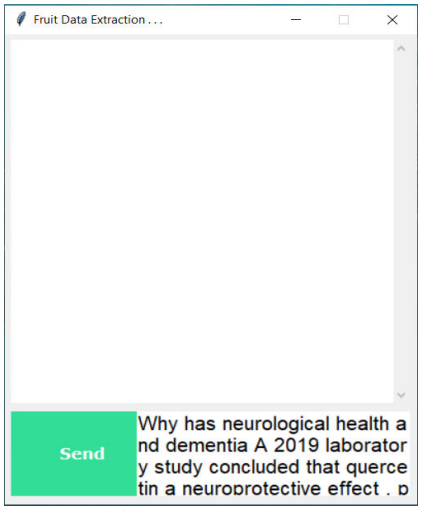

Fig 7: QA System

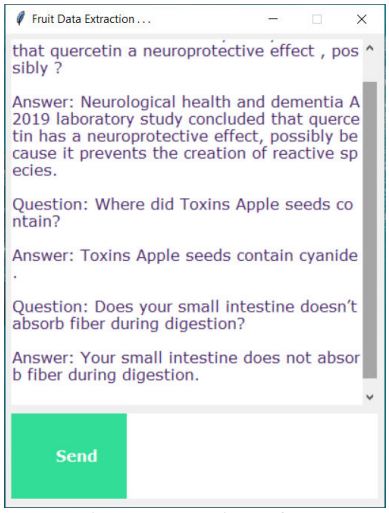

Fig 8: Extraction of QA

VI. CONCLUSION

Modern technology and a large amount of data have made it difficult for formers to receive the exact information online in the required time. There is various question answering systems developed but very few provide correct and efficient answers.

The Lack of a domain-specific QA system for unstructured data motivated our work for an Agriculture domain-specific QA system. A well-developed Question Answering system has to fetch the user query in the form of natural language and using an efficient classifier it has to classify user questions to give a correct and precise answer.

In this paper, we have studied various approaches and identified drawbacks. Hence, we can suggest implementing RNN (Recurrent Neural Network) approach. This technique contains internal memory and can remember the previous input to predict the next output. Due to its efficiency farmer queries can be answered accurately.

REFERENCES

1. Ajitkumar M, Pundge, Khillare S.A, C. Namrata Mahender. "Question Answering System, Approaches and Techniques: A Review". International Journal of Computer Applications (0975 – 8887) Volume 141 – No.3, May 2016
2. Sharvari Gaikwad, Rohan Asodekar, Sunny Gadia, Vahida Z. Attar. "AGRI-QAS Question-Answering System for Agriculture Domain". International Conference on Advances in Computing, Communication and Informatics (ICACCI) pp 1474-1478 2017.
3. Payal Biswas, Aditi Sharan, Nidhi Malik. "A Framework for Restricted Domain Question Answering System". 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908029 |

4. Manmita Devi, Mohit Dua, "ADANS:An Agriculture Domain Question Answeing System using Ontologies". Internaltional Conference on Computing, Communication and Automation (ICCCA) pp 122-127 2017
5. Waheeb Ahmed, Babu Anto P. "Arabic Question Answering System Based on Data Mining". International Journal of Scientific & Technology Research Volume 6, Issue 02, February 2017
6. Waheeb Ahmed, Babu Anto P. "Answer Extraction and Passage Retrieval for Question Answering Systems". International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 12, Dec 2016
7. Raju Barskar , Gulfishan Firdose Ahmed, Nepal Barskar. "An Approach for Extracting Exact Answers to Question Answering (QA) System for English Sentences". International Conference on Communication Technology and System Design 2011
8. Lokesh Kumar Sharma and Namita Mittal. "Answer Extraction in Question Answering using Structure Features and Dependency Principles" 9 Oct 2018. https://www.researchgate.net/publication/ 328189391
9. Amaral, C., Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C. and Vidal, D. 2007. Priberam's question answering system in QA@CLEF 2007. Springer, 5152, 364-371
10. Attardi, G. and Ciaramita, M. 2007. Tree Revision Learning for Dependency Parsing. In HLT-NAACL, Rochester, 388-395.
11. Azzopardi, L., Baillie, M. and Ruthven, I. 2007. Persuasive, Authorative and Topical Answers For Complex Question Answering. In TREC 2007, NIST, Gaithersburg, USA.
12. Babych, B. and Hartley, T. 2004. Extending the BLEU MT Evaluation Method with. Frequency Weightings, In Association for Computational Linguistics ACL '04, 621-628, Barcelona .
13. Banerjee, P. and Han, H. 2009. Drexel at TREC 2007: Question Answering. In International Florida Artificial Intelligence Research Society Conference (FLAIRS), Sanibel Island, Florida, USA.
14. Bloehdorn, S., Cimiano, P., Duke, A., Haase, P., Heizmann, J., Thurlow, I. and Völker, J. 2007. Ontology-Based Question Answering for Digital Libraries, In 11th European Conference on Digital Libraries (ECDL 2007), 14-25, Budapest, Hungary.
15. Bondarionok, A., Bobkov, A., Sudanova, L., Mazur, P. and Samuseva, T. 2007. Intellexer Question Answering. In Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA.
16. Bouma, G., Kloosterman, G., Mur, J., van Noord, G., van der Plas, L. and Tiedemann, J. 2007. Question Answering with Joost at CLEF 2007. In CLEF 2007 ,257-260.
17. Burger, J., Cardie, C., … Chaudhri, V. 2001.Issues, tasks and program structures to roadmap research in question and answering (QandA). Technical Report, NIST, Gaithersburg, USA.
18. Christie, J. 1999. Assessment of Essay Marking - focus on Style and Content. In 3rd International Computer Assisted Assessment Conference (CAA) , 39-45.
19. Chung, Y.-M., Pottenger, W. M. and Schatz B. R. 1998. Automatic Subject Indexing Using an Associative Neural Network. In 3rd ACM International Conference on Digital Libraries (DL'98), 59-68.
20. Cimiano, P., Haase, P., Sure, Y., Völker, J. and Wang, Y. 2006. Question answering on top of the BT digital library, In 15th international conference on World Wide Web, 861-862. Edinburgh, Scotland, UK.
21. Covington, M. A. 2001. A Fundamental Algorithm for Dependency Parsing. In 39th Annual ACM Southeast Conference, 95-102.
22. Dang, H. T., Kelly, D. and Lin, J., 2007 Question Answering Track. In TREC 2007, NIST, Gaithersburg, USA.
23. de Marneffe, M.-C. and Manning C. D. 2008. Stanford typed dependencies manual. Technical report, Stanford University.
24. de Marneffe, M.-C., Mac Cartney, B. and Manning, C. D. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In 5th Int. Conf. on Language Resources and Evaluation (LREC 2006), 449-454. Genoa, Italy,
25. Diekema, A.R., Yilmazel, O. and Liddy, E.D. 2004. Evaluation of Restricted Domain Question-Answering Systems. In ACL2004 Workshop on Question Answering in Restricted Domain, 2-7.
26. Enrique, A., Rosa, M. C., Manuel, F., Alvaro, O., Diana, P. and Pilar, R. 2005. Authoring of Adaptive computer Assisted Assessment of Free-text Answers, J. of Educational Technology and Society. 8(3): 53-65.
27. Fundel, K., Küffner, R. and Zimmer, R. 2007. RelEx – Relation extraction using dependency parse trees. J. Bioinformatics. 23(3): 365-371.
28. Gangolly, J. and Wu, Y.F. 2000. On the automatic classification of accounting concepts: Preliminary results of the statistical analysis of term-document frequencies, New Review of Applied Exert Systems and Emerging Technologies, 6, 81 – 88.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details