ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, *January 2015*

Rule-Knowledge Based Algorithm for Event Extraction

Prashant G. Desai [1], Sarojadevi H.[2], Niranjan N. Chiplunkar [3]

Faculty, Computer Science and Engineering Department, N.R.A.M.P., Nitte, Karnataka, India [1]

Professor & Head, Computer Science and Engineering Department, N.M.A.M.Institute of Technology, Nitte, Karnataka, India [2]

Principal, N.M.A.M. Institute of Technology, Nitte, Karnataka, India [3]

Abstract: The amount of electronic data being produced and communicated online is rapidly increasing. Most of these electronic documents contain lot of important & interesting information in unstructured format. Various approaches are used in the literature to automate the process of event extraction so as to conserve time and effort. This paper proposes an algorithm to automate the task of information extraction. We demonstrate that machine learning can be used for extracting event from the unstructured text such as dates places and subject of interest.

Keywords: Automation, Machine Learning, Unstructured Text, Information Extraction, Natural Language.

I. Introduction

Since a decade or so, there has been an exponential increase, in the information available on the internet. Most of such information is available in the form of natural language documents. Users often find it difficult to locate needed information in these documents. Therefore, technology that can extract most important information available in the documents has gained significance, and document summarizing technology has become an extremely important area of research. One way of providing information as per the user's requirement is information extraction [10]. Information extraction is the task of locating specific pieces of data from a natural language document.

There are many domains for which information extraction can be applied: it is extremely helpful in medical field to study the symptoms of diseases, keep track of the various health parameters of the patients; to analyse the sentiments of the blogs posted on social media; to get the highlights from the news articles; to obtain the product reviews and many more. This paper discusses its use in documents which contain information regarding seminar, conference announcements. Event extraction has become one of the most important tasks of information extraction, since it is the key to many applications in natural language processing such as personalized news systems, question answering and document summarization [11]. The process of discovering useful information from event documents is termed "event extraction" for which text mining approaches are used. Frederic et. al. [2] distinguish between three main approaches to event extraction.

A. *Data driven event extraction*

Data-driven text mining approaches are based on quantitative method, probability theory and require large text corpora to develop the model. However, this approach lacks to consider semantics of the text.

B. *Knowledge driven data extraction*

Knowledge-driven text mining is based on patterns that express rules representing expert knowledge & considers meaning of the text. Predefined or discovered patterns are used to extract the events.

C. *Hybrid event extraction*

As both approaches have their disadvantages, in practice combining two methods could yield the best results.

II. RELATED WORK

Lot of research has been done in the field of event extraction, text mining, question answering, and text summarization. In this section we discuss work related to our research. Dang Truong Son et. al. [3] propose an approach by exploiting a large number of available pairs of question-answer documents in order to search the best similar question to user's question. This approach involves converting the given question by expanding the sentence with synonyms and trimming the list of extended sentences by language modeling, thereby moving the user's question to suit the question-answer database. The process of finding the most relevant question and returning the corresponding answer will be made by combining the techniques of natural language processing and information retrieval.

The question processing system described in [2] involves finding the syntactic structure of a question. Questions with the same syntactic structure can be treated the same and their differences handled by later processing. These similar questions are processed with the same regular expression and their syntactic structure determines their properties. Another addition to the system is further processing of the entities and relations. Adjectives and extra information are separated from the main part of entities to allow simpler searching and provide additional data.[12] Explores the task of creating a timeline for historical Wikipedia articles, such as those describing

DOI 10.17148/IJARCCE.2015.4116

wars, battles, and invasions. Authors focus on extracting only the major events from the article, particularly those associated with an absolute date. For determining important events, authors take the assistance of the EVITA [12] program on the article, which labels all possible events in the TimeML format. It places XML “EVENT” tags around single words defining events. A website is maintained where a user can view already processed Wikipedia articles or request the processing of new ones. The website allows users to view several aspects of the article.

The research described by the authors of [6] focus on modelling the content of news events by their semantic relations with other events, and generating structured summarization. This research proposal is aimed at structured summarization for news topics in order to help the users answer “What, When, Why, Where, How” questions. Extending from traditional summarization, researchers consider various relational types such as causal, temporal, spatial and topic hierarchical relations since they are intuitively close to typical “Wh-questions" that people likely to ask when they are looking for information about the news topics.

The basic algorithm steps proposed in the work of [13] are illustrated below:

1. Input: A Query as a Sentence.

2 Convert this sentence into predicate logic after then clause form.

3. Processing and semantics analysis of the sentence with the help of predicate logic and clause form. Now question is to be asked and converted into predicate logic and then clause form. 4. Construct the rules or knowledge base using resolution and unification algorithm. Finally the clause form of the question is resolved to final answer using resolution and unification algorithm.

5. Similarity comparisons between input sentence and database of the text by using algorithm.

6. Finally relevant answer is retrieved with respect to corresponding query sentence.

7. Repeat step 1 to 7 for another query sentence.

8. End

Authors have developed a Question Answering (QA) System for English sentences. The user should be able to access answer of their questions in a user friendly way, that is by questioning the system from the given English paragraph and the system will return the intended answer by searching in context of the paragraph using the repository of English dictionary. In this paper a Question/Answering system that takes advantage from category information by exploiting several models of question and answer categorization is presented. The focus is on context based retrieval of information. This paper provides a novel and efficient method for extracting exact textual answers from the returned documents that are retrieved by traditional IR system in large-scale collection of texts.

The work presented by [1] is aimed to reveal the implicit knowledge present in news streams. This knowledge is expressed as a hierarchy of topic/subtopics, where each topic contains the set of documents that are related to it and a summary extracted from these documents. Summaries so built are useful to browse and select topics of interest from the generated hierarchies. Researcher of [8] proposes an annotation scheme to cover different types of causality between events, techniques for extracting such relations and an investigation into the connection between temporal and causal relations. There is a special focus on the relation, because causality is presumed to have a temporal constraint. Here two types of event relations to be extracted from text are considered, which are temporal relations and causal relations.

[9] discusses parameters for text quality measurement. Listed below are few text quality aspects.

Grammaticality – the text should not contain non-textual items (i.e., markers) or punctuation errors or incorrect words.

Non-redundancy – the text should not contain redundant information reference.

Clarity – the nouns and pronouns should be clearly referred to in the summary. For example, the pronoun he has to mean somebody in the context of the summary. Coherence and structure – the summary should have good structure and the sentences should be consistent. Many human language technology (HLT) tasks such as Information extraction are traditionally evaluated using Precision, Recall and F-measure [4]. We have evaluated results obtained based on these three metrics.

III.SYSTEM ARCHITECTURE

The proposed work is shown in the figure 1. This proposed machine learning algorithm consists of following modules.

A. File classifier

The work is implemented in such a way that user can provide Ms-Word or pdf format of input document. This module identifies whether it is a MS- Word or pdf document.

B. Communication with the external system

This module is responsible for communicating with the external tools used while implementing our work. We use mainly two tools PDFOne [7] and Apache POI [8].

1) PDFOne

Gnostice PDFOne is a versatile PDF SDK for implementing PDF-related features in Java applications. PDFOne can create, edit, view, print, encrypt, decrypt, merge, split, reorganize, bookmark, annotate, watermark, and stamp PDF documents. The API hides the complexity of the PDF format and enables to quickly implement sophisticated PDF features. PDFOne is entirely written in Java code.

2) Apache Poor Obfuscation Implementation (POI)

This is a Java API to Handle Microsoft Word Files. HWPF is the port name of Microsoft Word 97(-2007) file format to Java. Word document can be considered as very long single text buffer. HWPF API provides "pointers" to document parts, like sections, paragraphs and character runs. A Word file is made up of the document text and data structures containing formatting information about the text. The entry point for HWPF's reading of a Word

DOI 10.17148/IJARCCE.2015.4116

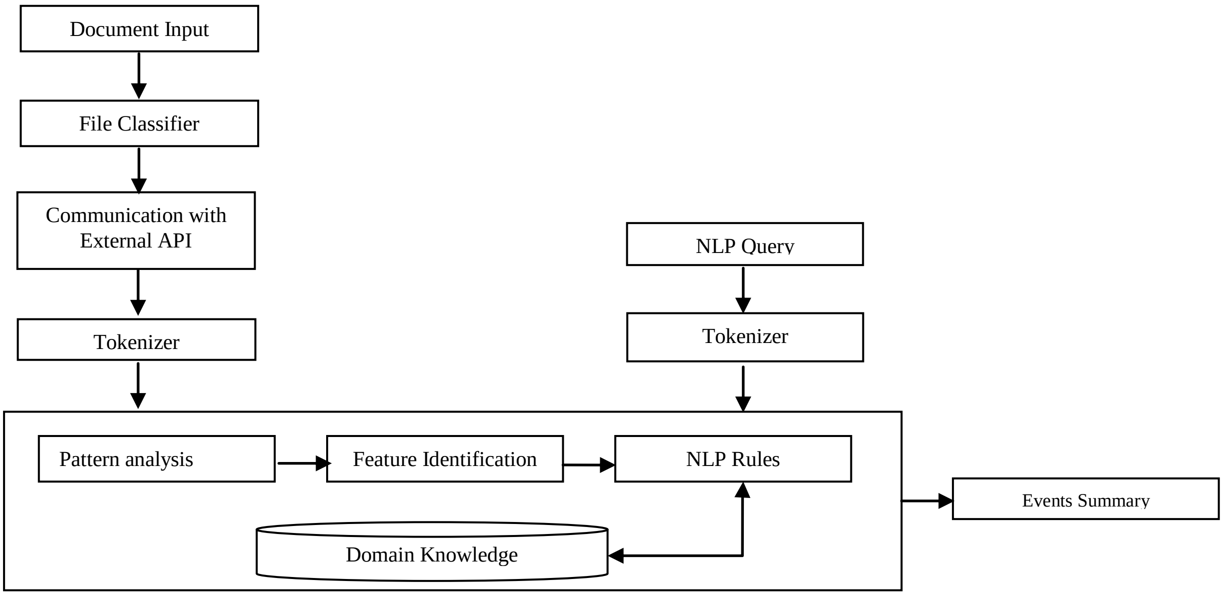


Fig. 1 System architecture

file is the File Information Block (FIB). This structure is the entry point for the locations and size of a document's text and data structures. The FIB is located at the beginning of the main stream.

C. Tokenizer

Tokenization is a process of breaking input stream of characters into an ordered sequence of words like units, usually called tokens that can be used for further processing [13]. These tokens may correspond to words, numbers, and punctuation marks. Tokenization is a prerequisite in order to perform any Information extraction task.

D. Pattern analysis

First this module identifies the boundary of sentence and then breaks the text of whole document into sentences. Then each sentence is broken down into words. Next step is to identify dates. Identification of dates is done by processing every word in each sentence. Presently, our work is focused on extracting important dates from the event documents such as date of registration, paper submission date, date of acceptance etc.

Rules used by the algorithm for learning dates:

1) Rule 1

2 digit Number Separator 2 digit Number Separator 4 Digit Number

2) Rule 2

2 digit Number Separator 2 digit Number Separator 2 Digit Number

3) Rule 3

2 digit Number Separator First Three Letters of Month Separator 4 Digit Number

4) Rule 4

2 digit Number Separator First Three Letters of Month Separator 2 Digit Number

Located : VBN

5) Rule 5

2 digits Number Separator Month Separator 4 Digit Number

6) Rule 6

2 digits Number Separator Month Separator 2 Digit Number

These rules are then incorporated in the algorithm as below:

```
1. Break text from document into lines.
2.For each line of text in the document
If apply rule -1 to rule -6 to verify dates are present then
        Output current line of text
    End if
End for
```

E. Feature identification

In this step feature of the tokens such as Noun, Verb, Adjective, Adverb, Pronoun, Proper noun, etc are identified. This process of identifying the feature of tokens is also called as Part-Of-Speech (POS) tagging. The algorithm takes the assistance of Stanford POS tagger [14] for assigning tags to each of the tokens. These features of tokens are then used to identify the places.

F. NLP Rules

This module performs to tasks: Identifying places and identifying tracks from the event document.

G. Identifying places

This module first generates candidate sentences which possibly contain places. This is accomplished by filtering the sentences having NNP words. It is because the venues are generally presented as phrases like “located in Coimbatore, Tamil Nadu, India”. After tagging, the phrase would look like below:

DOI 10.17148/IJARCCE.2015.4116

In: IN
Coimbatore:NNP
Tamil : NNP
Nadu : NNP
, : ,
India : NNP

Now, following rules are formulated to process these candidate sentences

1) *Rule7*

The count of NNP words in each candidate sentence must be at least 2

2) *Rule 8*

If a tag "," is present then Tag for the previous word & Tag for the next word must be NNP. Such phrases are picked as name of the places.

3) *Rule 9*

If the system is unable to decide on the place then the assistance of geonames.org is taken online to finalize the place name.

H. *Identifying tracks*

Tracks are the areas in which research articles are accepted in a conference. When such areas or subjects are mentioned in the event document, they are made of multiple technical words. Therefore, to identify tracks it is very much essential to know whether a word is a technical word. In order to decide whether word is technical the module is assisted by the knowledge base. This knowledge base is a collection of all possible technical words with collection of over 1500 technical words.

Outline of the algorithm used to identify technical words in a sentence are as follows:

```
1. Tokenize the sentence into words.
2 For each word in the sentence
        Check if the word is a technical word
        If the word is technical then
                Retain the word
        Else
                Just discard the word
        End if
End for
3. For each retained words in each sentence in the previous step -Check if multiple technical words appear together and if this is true add such collection to list of tracks
 End for
```

I. *Query based Dialogue Management*

This module allows making interaction with the user for the required dialogue. Presently we are concentrating on key words based dialogue interface. This module is responsible for tokenizing and tagging each word in the question with an appropriate part-of-speech tag. Then *stop words* are removed from the question sentence. Remaining words are treated as keywords or index words. Response is extracted based on the index terms. For instance, if the "date" is listed as an index or key term, then all available dates with respect to other key terms are extracted.

The dialogue management control interface is presently focusing on WH-questions. Few of the possible WH-questions asked and response of the system for these questions are listed below:

1) *Q: Which is the venue of the conference?*

System: Coimbatore, Tamil Nadu, India

2) *Q: What are the important dates in the conference?*

System: 30 & 31 January 2013
November 30, 2012: Full Paper Submission
December 15, 2012: Paper Acceptance Notification to Authors
December 20, 2012: Registration, Copyright Submission, Final Camera Ready Paper Due
January 30 & 31, 2013: Conference

3) *Q: What are the areas of research in which papers can be submitted?*

System:
platform computer science
applications computer
Computer Applications
Computer Networks
Concurrency Parallelism
Domain Languages
Secure Computing
Mobile Networks
Cloud Computing
Social Networking
Distributed Computing
Semantic Web
Wireless Sensor Networks
Service Oriented Architecture

IV. EVALUATION METRICS

Some of the popularly used evaluation metrics for text mining approaches are Precision, Recall, and F-Measure.

A. *Precision*

Precision measures the number of correctly identified items as a percentage of the number of items identified. In other words, it measures how many of the items that the system identified were actually correct, regardless of whether it also failed to retrieve correct items. The higher the Precision, the better the system is at ensuring that what has been identified is correct. It is formally defined as

$$\text{Precision (P)} = \frac{\left(\text{Correct} + \frac{1}{2}\text{Partial}\right)}{\left(\text{Correct} + \text{Spurious} + \text{Partial}\right)}$$

B. *Recall*

Recall measures the number of correctly identified items as a percentage of the total number of correct items. In other words, it measures how many of the items that should have been identified actually were identified, regardless of how much spurious identification was made. The higher the Recall rate, the better the system is at not missing correct items. Recall is formally defined as

DOI 10.17148/IJARCCE.2015.4116

$$\text{Recall(R)} = \frac{\left(\text{Correct} + \frac{1}{2}\text{Partial}\right)}{\left(\text{Correct} + \text{Missing} + \text{Partial}\right)}$$

C. F-measure

The F-measure is often used in conjunction with Precision and Recall, as a weighted average of the two. If the weight is set to 0.5 (which is usually the case), Precision and Recall are deemed equally important. F-measure is formally defined as

$$\text{F} - \text{measure} = \frac{\left(\beta^2 + 1\right)\left(\text{P} * \text{R}\right)}{\left(\beta^2\text{R} + \text{P}\right)}$$

Where β reflects the weighting of P vs. R. If P and R are to be given equal weights, then we can use the equation

$$\text{F1}= \frac{\left(P*R\right)}{0.5*\left(\text{P}+\text{R}\right)}$$

V. RESULTS AND ANALYSIS

This section describes performance and results obtained out of the system implemented. The system described is tested on ten event documents having different words count to extract intelligence from the text.

TABLE I
RESULT ANALYSIS

Length (Number of Words in the document)	Information Extracted	Correct	Partial	Spurious	Missing	Precision	Recall	F Measure
149	Date	4.00	0.00	0.00	0.00	1.00	1.00	0.25
	Place	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Subjects of Interest	Na	Na	Na	Na	Na	Na	Na
167	Date	2.00	0.00	0.00	0.00	1.00	1.00	0.25
	Place	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Subjects of Interest	Na	Na	Na	Na	Na	Na	Na
187	Date	Na	Na	Na	Na	Na	Na	Na
	Place	2.00	0.00	0.00	0.00	1.00	1.00	0.25
	Subjects of Interest	31.00	5.00	2.00	6.00	0.88	0.80	0.21
189	Date	3.00	0.00	0.00	0.00	1.00	1.00	0.25
	Place	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Subjects of Interest	Na	Na	Na	Na	Na	Na	Na
212	Date	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Place	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Subjects of Interest	Na	Na	Na	Na	Na	Na	Na
296	Date	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Place	Na	Na	Na	Na	Na	Na	Na
	Subjects of Interest	Na	Na	Na	Na	Na	Na	Na
348	Date	3.00	2.00	0.00	0.00	0.80	0.80	0.20
	Place	1.00	1.00	0.00	0.00	0.75	0.75	0.19
	Subjects of Interest	11.00	0.00	2.00	9.00	0.85	0.55	0.17
352	Date	5.00	0.00	0.00	0.00	1.00	1.00	0.25
	Place	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Subjects of Interest	Na	Na	Na	Na	Na	Na	Na
367	Date	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Place	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Subjects of Interest	Na	Na	Na	Na	Na	Na	Na
967	Date	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Place	1.00	0.00	0.00	0.00	1.00	1.00	0.25
	Subjects of Interest	Na	Na	Na	Na	Na	Na	Na

DOI 10.17148/IJARCCE.2015.4116

A. Experimental setup

Initially a file is input to the system. Based on the file extension the file classifier identifies whether the input file is a MS-Word document or "pdf" document. If it is MS-Word document then the file classifier feeds file into Poor Obfuscation Implementation File System (POIFS) subcomponent of Apache tool otherwise file is transferred to PDFOne component for reading "pdf" document.

B. 5.2 Processing a Word File

1. An instance of POIFSFileSystem is created using an input stream from which data is to be read.

2. The input stream is read till EOF.
3.The constructor HWPFDocument which uses POIFSFileSystem containing the Word document as parameter is used for loading the Word document.
4. Text from the loaded word document in step-3 is extracted as paragraphs using constructor WordExtractor for further processing.

C. Processing a pdf file

1. An instance of PDFOne class is used to create a NEW pdf document from scratch and load an existing document.
2. The lines of text from the loaded document are then extracted for further processing.

D. Input documents

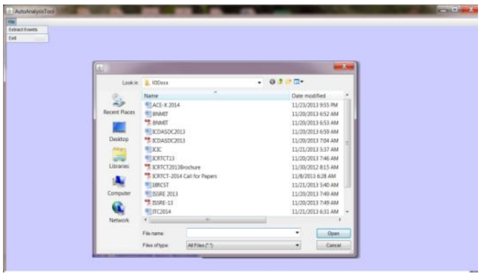
We have taken 10 complex event documents belonging to both national and international journals and conferences within India and abroad as case study. The documents chosen are randomized in such a way that, few contain only dates, few contain only places, few contain only subjects of interest and few contain combination of this information. We then evaluate the obtained results with respect to the performance parameters discussed in section *IV*.

E. Result analysis

Results are obtained, analysed and presented in Table I.
The dates embedded within the unstructured text are identified to the near perfection. Machine learning approach is used for identifying the places. It is concluded from the earlier researches in the field of text mining that 100% accuracy cannot be achieved from the machine learning. Results are encouraging when we applied machine learning for extracting names of the places. Knowledge based rules are employed for extracting subjects of interest.

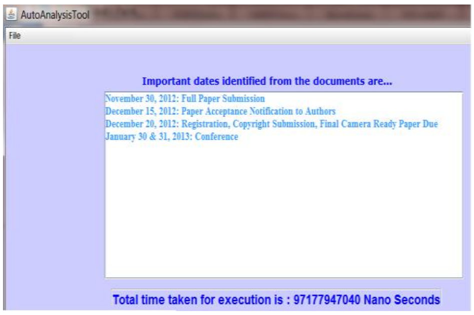
F. Working of the system

The interactive dialogue management part of the software developed for the text processing is shown in figure 2 through figure 5.



Name	Date modified
ACE-X 2014	11/23/2013 9:55 PM
BNMIT	11/20/2013 6:52 AM
BNMIT	11/20/2013 6:53 AM
ICDASDC2013	11/20/2013 6:59 AM
ICDASDC2013	11/20/2013 7:04 AM
ICIC	11/21/2013 5:37 AM
ICRTCT13	11/20/2013 7:46 AM
ICRTCT2013Brochure	11/30/2012 8:15 AM
ICRTCT-2014 Call for Papers	11/8/2013 6:28 AM
IJIRCST	11/21/2013 5:40 AM
ISSRE 2013	11/20/2013 7:49 AM
ISSRE-13	11/20/2013 7:49 AM
ITC2014	11/21/2013 6:31 AM

Fig. 1. Input file selection

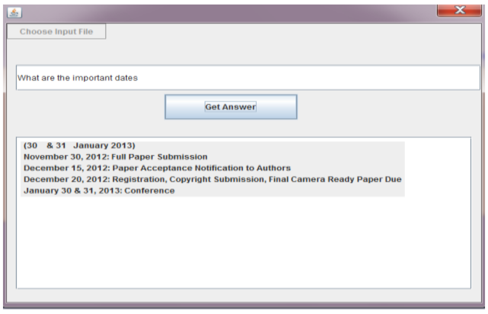


Important dates identified from the documents are...

November 30, 2012: Full Paper Submission
December 15, 2012: Paper Acceptance Notification to Authors
December 20, 2012: Registration, Copyright Submission, Final Camera Ready Paper Due
January 30 & 31, 2013: Conference

Total time taken for execution is : 97177947040 Nano Seconds

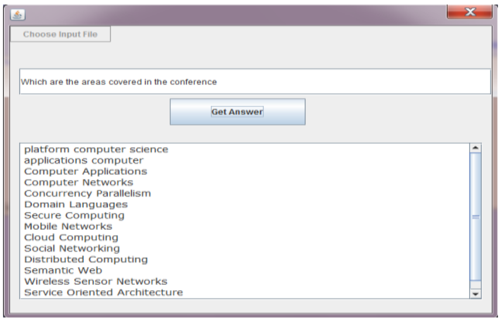
Fig.3. Extracted information – important dates



What are the important dates

(30 & 31 January 2013)
November 30, 2012: Full Paper Submission
December 15, 2012: Paper Acceptance Notification to Authors
December 20, 2012: Registration, Copyright Submission, Final Camera Ready Paper Due
January 30 & 31, 2013: Conference

Fig 4. Extracted information against Dialogue- Important Dates



Which are the areas covered in the conference

platform computer science
applications computer
Computer Applications
Computer Networks
Concurrency Parallelism
Domain Languages
Secure Computing
Mobile Networks
Cloud Computing
Social Networking
Distributed Computing
Semantic Web
Wireless Sensor Networks
Service Oriented Architecture

Figure 5: Extracted information against a Dialogue - Area of interest in a Conference

VI. CONCLUSION

This paper presents a new approach for extracting information from event documents. A new rule-knowledge based algorithm is developed for the information extraction. Results obtained confirm that the dialogue management interface can process the queries presented in the form of natural language. This dialogue management interface is based on domain specific knowledge base and key word matching concepts.

DOI 10.17148/IJARCCE.2015.4116

REFERENCES

[1] Aurora Pons Porrata , Rafael Berlanga Llavori , Jose Ruiz Shulcloper, “Topic Discovery Based on Text Mining Techniques”, Proceedings of Information Processing and Management-43, 2007, pp 752–768.

[2] Chris Whidden “Simple and Effective Question Processing Using Regular Expressions and Word Net”, 2005.

[3] Dang Truong Son, Dao Tien Dung, “Apply a Mapping Question Approach in Building the Question Answering System for Vietnamese Language ”, Proceedings of JETE, 2012, pp380-386.

[4] Dayana Maynard, Wim Peters, Yaoyong Li, “Metrics for Evaluation of Ontology Based Information Extraction”, Proceedings of WWW workshop, 2006.

[5] Frederick Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong, “An Overview of Event Extraction From text”, Proceedings of 1st workshop on detection, representation & exploration of events, Proceedings of ISWC, 2011.

[6] Giang Binh Tran, “Structured Summarization for News Events”, Proceedings of IW3C2, 2013

[7] “http://www.gnostice.com/PDFOne_Java.asp”

[8] “http://poi.apache.org/hwpf/index.html”

[9] Josef Steinberger, Karel Jezek, “Evaluation Measures for Text Summarization”, Proceedings of computing and informatics, Vol. 28, 2009

[10] Mary Elaine Calif, “Relational Learning Techniques for Natural Language Information Extraction” , Ph.D. Dissertation, 1998

[11] Paramita Mirza, “Extracting Temporal and Causal Relations between Events”, Proceedings of the ACL 2014 Student Research Workshop, 2014, pp 10–17

[12] Rachel Chasin, “Event and Temporal Information Extraction towards Timelines of Wikipedia Articles”, UCCS REU, 2010

[13] Raju Barskar, Gulfishan Firdose Ahmed, Nepal Barska, “An Approach for Extracting Exact Answers to Question Answering (QA) System for English Sentences”, Proceedings of ICCTSD, 2012, pp 1187 – 1194

[14] Stanford University POS Tagger

DOI 10.17148/IJARCCE.2015.4116