ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

TEXT INFORMATION EXTRACTION USING RULE BASED METHOD

Kavita Namdev*, Ankit Patidar, Abhishek Patel

ABSTRACT

Information is hidden in large volume of files thus it is necessary to find useful information and extract it from file contents. Information Extraction (IE) is the task of automatically extracting structured information from unstructured or semi-structured documents. The data in all available files out of total 80% falls in category of unstructured text or semi structured text, this data is typically heavy, but may contain facts as well as very useful information. When we search any useful information from files is very tedious, since searching algorithms have high complexity and require time to search each word. Or in today's Era everything is going to be store in form of files in computers and both online and offline sources generate large amount of text data on daily basis. So gathering or retrieval of information from large volume of data via searching algorithm is not preferred so we use concept of information extraction. Many methods have been proposed for automating the process of extraction, but due to the heterogeneity and lack of structure of file contents automated discovery of information still faces many challenges in new researches. This research paper will going to presents a system which is a powerful toolkit for rule-based information extraction. Developed system is based on top down approach of rule based method and provides versatile information processing and advanced extraction techniques. We thoroughly describe the system and its capabilities for extraction and performance calculation based n certain parameters.

KEYWORDS: Information Extraction, Text routing, Text Mining, knowledge discovery, structured data, Semi-Structured data.

INTRODUCTION

The field of automatic IE has only been driven forward in last decade. Two factors have been important for the development of the field. Firstly exponential growth in amount of both online and offline textual data and second focus of field through MUC during last 10 years. Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources. IE systems have undergone a continuous and interesting development and recently, many extended works have been suggested and developed. The core idea behind IE systems is the aim to retrieve specific and desired information from documents of natural language text. These extraction processes are conducted automatically using computer methods, and this process has developed rapidly within the timeframe because of the development of NLP tools and techniques. IE systems have been developed to extract information from different types of text; structured text, semi-structured and free text. Recently, this method has been extended to include the extraction of information from images and videos [17].

- Unstructured Data(Free-text): This kind of data neither have a pre-defined data model nor organized in a pre-defined manner which include books , articles, E-Mail, etc.
- Semi-structured Data: The data that is presented and formatted in a high quality manner in a specific domain for instance, information about the economy, education, medicine and so forth.
- Structured Data: This type of document is highly structured, organized, and well formatted. A key example is databases, records, spreadsheets etc.

The main goal of information extraction is to extract specific structured information from the unstructured or semi-structured text, and use these information more accessible to people and also making information more machine-process able, and Put information in a semantically precise form that allows further inferences to be made by computer algorithms.

IE has application in a wide range of domains which includes biomedical researchers in gene. Proteins and other entities, financial analysis in which Professionals seek specific pieces of Information from news articles to help day

to day marketing and in web technology as search engines have became an integral part of daily life of common people.IE also help in such area as entity search, Structured search and question answering etc.

As an example of text Information Extraction, we can describe a system that processes a text file and extract information regarding name, Date of Birth and location of a person who have been introduce in a given text file. This system will attempt to extract information matching given parameters and ignore all other data.

Consider the following English sentence: "In 2015, James Smith and Von Hall founded Techsoft.Ltd" We can extract information like, Founderof(James Smith, Techsoft.Ltd), Founderof(Von Hall, Techsoft.Ltd), foundedIn(Techsoft.Ltd, 2015)

IE is an important research area and in this field. There are many methods emerged which includes mainly – Dictionary based method, Rule based method and wrapper induction. The rule based method is different from others as they have several general rules to extract information from text. Rule based it is an important approach for processing the increasingly available amount of the unstructured data. The manual creation of rule based applications is a time consuming and tedious task. This required qualified knowledge engineers. The cost of this process can be reduced by providing a suitable rule language and extensive testing support. Rule based IE consist of a specification of a text base rule language and an interpreter which is able to apply the rules on documents in order to identify new information. It has been concluded that the rule based approach has been proven capable of delivering reliable IE, with extremely high accuracy and coverage results. However this approach requires an extensive time consuming and manual study of word classes and phrases.

Applications

- News Tracking: A classical application of information extraction, which has spurred a lot of the early research in the NLP community, is automatically tracking specific event types from news sources. The popular MUC and ACE competitions are based on the extraction of structured entities like people and company names, and relations such as "is-CEO-of" between them [2].
- Biomedical Researches:- They are often need to sift through a large amount of scientific publications to look for discoveries related to particular genes, proteins or other biomedical entities. To asset this afford, simple search based on keyword matching may not suffice because biomedical often have synonyms and ambiguous names.
- Financial professionals: - They also required information extraction to seek specific pieces of information for news articles to help their day-to-day decision making.
- Community Websites: Another example of the creation of structured databases from web documents is community web sites such as DBLife and Rexa4 that tracks information about researchers, conferences, talks, projects, and events relevant to a specific community. The creation of such structured databases requires many extraction steps: locating talk announcements from department pages, extracting names of speakers and titles from them, extracting structured records about a conference from a website and so on [2].
- Opinion Databases: There are innumerable web sites storing unmoderated opinions about a range of topics, including products, books, movies, people, and music. Many of the opinions are in free text form hidden behind Blogs, newsgroup posts, review sites, and so on. The value of these reviews can be greatly enhanced if organized along structured fields. For example, for products it might be useful to find out for each feature of the product, the prevalent polarity of opinion [2].

LITERATURE SURVEY

IE Methods

Information Extraction is an important research area and in this field this are many methods emerged which mainly includes – Dictionary based method, Rule based method and wrapper induction. All these methodologies have immediate real-life applications. Information extraction has been applied, for instance, to part-of-speech tagging [6], named entity recognition [5], shallow parsing [7], table extraction [9], and contact information extraction[8].

Dictionary based method

The Dictionary based method also known as pattern based systems which is a traditional information extraction approach in this systems first construct a pattern (template) dictionary, and then use the dictionary to extract needed information from the new untagged text.[1]

Rule based method
Different from the dictionary based method, the rule based method use several general rules instead of dictionary to extract information from text. The rule based systems have been mostly used in information extraction from semi-structured files[1]. There are a variety of approaches to constructing Rule Based IE systems. One approach is to manually develop information extraction rules by encoding patterns (e.g. regular expressions) that reliably identify the desired entities or relations. For example, the Suiseki system [4] extracts information on interacting proteins from biomedical text using manually developed patterns [3]

However, in our experience rule-based techniques provide a viable alternative especially since these allow for rapid-prototyping capabilities, that is, by starting with a minimal rule set that can be extended as needed [13].

A Rule Based system consists of a list of rule elements that are made up of three parts: The mandatory matching condition of a rule is given by a TypeExpr or a StringExpr and creates a connection to the document. Second the optional QuantifierPart defines greedy or reluctant repetitions of the rule element, similar to regular expressions. Then the third, additional conditions and actions in the ConditionActionPart add further requirements and consequences to the rule element.[13] Usually an information extraction system supports one of the two basic approaches of extraction, namely, Knowledge Engineering Approach and Automatic Training Approach.

Knowledge engineering approach
In order to extract information from available texts using a system which supports a knowledge engineering approach a set of extraction rules must be written manually. A person who creates such a type of system, or is responsible for writing those rules (i.e., a knowledge engineer) must be an expert in the knowledge domain chosen for extraction or at least must be closely familiar with it. Since this approach involves writing rules, in some sources it is called as a simple rule-based approach [17].

Automatic training approach
In this case there is no need to design extraction rules manually. Therefore a person who is responsible for the information extraction process does not have to know how to write rules and how a system works. A machine learning algorithm implemented in the information extraction system creates those rules. In order to do that the algorithm must have access to a large number of training texts related to the chosen domain. Those texts must be annotated manually in advance to provide examples on which the algorithm can learn and produce extraction rules. Thereby, the engineer must provide the set of training documents and be able to annotate them. Since one of the main goals of the project is writing a set of extraction rules for a specific domain the question which method to prefer does not arise [16].

Wrapper induction
Wrapper induction is another type of rule based method which is aimed at structured and semi-structured documents such as web pages. A wrapper is an extraction procedure, which consists of a set extraction rules and also program codes required to apply these rules. Wrapper induction is a technique for automatically learning the wrappers[1]. The typical wrapper systems include WIEN [12] , Stalker[11], and BWI [10]. which use the principle of wrapper induction. WIEN is the first wrapper induction system.

Performance evaluation criterion-
In information extraction and pattern recognition performance mainly relies on recall and precision value (possibly combined in a F- measure) to assess performance of extraction. Precision which is also called positive predictive value is defined as the fraction of extracted instances that are relevant, while recall which is also known as sensitivity is the fraction of relevant instances that are extracted. Both precision and recall are therefore based on an understanding and measure of relevance. Recall (R) is the proportion of class members that the system assigns to the class. Precision (P) is the proportion of members assigned to the class that really are class members. Fallout (Fal) computes the proportion of incorrect class members given the number of incorrect class members that the system could generate. Ideally, recall and precision are close to 1 and fallout is close to 0.
Precision Rate = ard value /ad value

Recall Rate = ard value / trd value
Where:
ard value = number of relevant documents in the result list.
trd value = total number f relevant documents in the document base.
ad value = number of documents in the result list.

F-measure : When comparing two classifiers, it is desirable to have a single measure of effectiveness. The F-measure, derived from the E-measure is a commonly used metric for combining recall and precision values in one metric. It is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score
.F-measure=2*(precision*recall) / (precision+ recall)

Related Work
Since then several architectures have been developed to facilitate the process of the information systems development by providing the common platform for systems' components design, integration and reuse. Among them are the Unstructured Information Management Architecture (UIMA), the General Architecture for Text Engineering (GATE), the Architecture and Tools for Linguistic Analysis Systems (ATLAS), the Automated Linguistic Processing Environment (ALPE). AutoSlog, LIEP system, PALKA system, CRYSTAL are some of the learning system that generates extraction rules [18].

Information extraction automation has become more popular due to some restrictions of the previous approach, like time and effort consumption. Among the automated systems are WHISK, RAPIER, WIEN, SRV (supervised); IEPAD, OLERA (semi-supervised); DeLa, RoadDunner, DEPTA (unsupervised). Five of the most common supervised learning techniques are the Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM), Conditional Random Fields (CRF), Sup-port Vector Machines and Decision Trees [18]. There are many such systems are developed before on the basis of various predefined IE methds.

PROBLEM DEFINITION

There is tremendous amount of textual data present in both online and offline sources, so there is growing need for systems that extract information automatically from text documents One type of IE, named entity recognition, involves identifying references to particular kinds of objects such as names of people, companies, and locations [16].An enormous amount of information exist in natural language form. If these information is to be automatically manipulated and analyzed it first must be distilled into a more structured form in which the individual facts are accessible. The text document in file present in many formats which includes unstructured text, semi-structured text and structured text. This results in the irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in a any structured format just like in fielded from in databases.

More recently it has been recognized that by setting a goal of selective information structuring means information extraction we can define a range of tasks that appears within reach of current technology. A mature information extraction technology would allow us to rapidly create extraction system for new task whose performance was on a par with human performance [14].

Let us consider a situation in which the whole summary of any topic is given in which the information present is too less and whose retrieval and extraction is tedious work so an information extraction system can serve as a front end for high precision information collection and gathering. This research works will mainly emphasis on the development of the "Rule Based Information Extraction System" with much more better performance results then the systems developed earlier.

PROPOSED SOLUTION

Architecture of developed Information Extraction System
To know the IE system correctly we must properly understand the each and every phase of proper extraction process in which the input data is passed from various sub steps of extraction and in the final step facts are integrated and the pertinent facts are translated into a desired required output format.

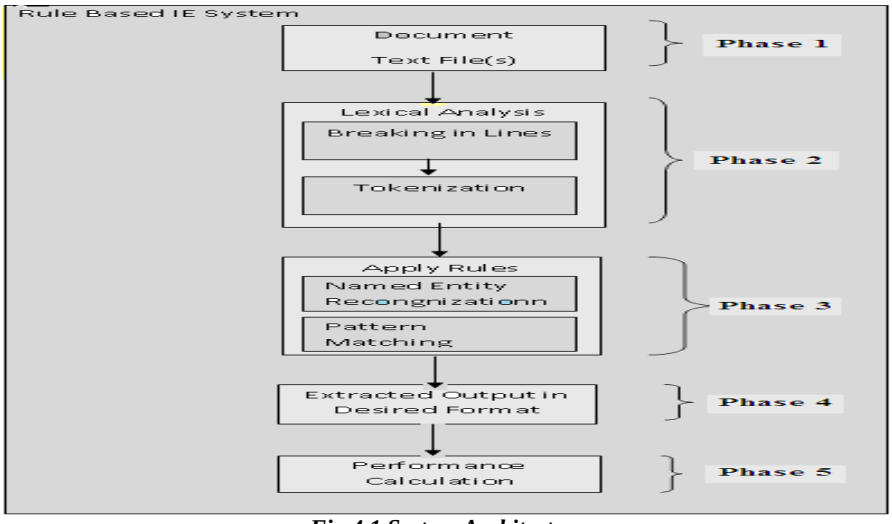

Fig 4.1 System Architecture

1st Phase:

This is a first phase of system which is considered as a input phase for a system which act as a front end for the user through which user can able to communicate with the system by following ways:

a) User is allowed to enter the document (file) on which the system acts and use these file for extraction of desired information from it.
b) Secondly in input phase user is also allowed to select parameter from the parameter list and this parameter acts as a basis for the task of extraction of information from the selected file.

2nd Phase:

This is the second phase of the system which referred as a lexical analyzer and it is used for the proper working of algorithm. This phase is further divided into two parts namely: line breaking and tokenization

a) Line breaking: In this part the input file document is braked up in to the sequence of lines and which act as a input for next part.
b) Tokenization: It is process of forming token(String) from the input line and token(also called lexeme) is defined as set of possible character sequence. Tokens are used in further phases for pattern matching and named entity recognition.

3rd phase:-

This is third phase of the system in which application of rules takes place and this is central and main part of the Rule Based Information Extraction where actually braked units are matched to the regular expressions which are defined already in system. This phase is further divided into two parts namely: Named entity recognition and pattern matching.

- NER(Named entity recognition): This part is also considered as entity identification or entity extraction part which is sub part of application phase and which is used to locate and classify the elements(lexical units) in the text into the predefined categories such as the name of person ,organization, location , date , currency and many more.
- Pattern matching: This part acts as checking a given sequence of tokens for the presence of constituents of some patterns. Sequence patterns are described in the system using in form “Regular Expression”. Regular Expression can be said as-

4th Phase:

This is the output phase of the system which is main phase for the end user. In this phase the extracted information is provided to the user and user can use these information in accordance to need. Either these extracted information may directly use or may be permanently store to database to give the extraction result the persistent storage which may refer by the user in future for further use as according to the requirement of the user.

5th Phase:
This phase concern with the performance of the developed IE system. In which the following parameter are used to determine the overall performance of the system. Which are: Recall Rate, Precision Rate and F-measure. In this phase the user is also allow to view various performance parameters in the graphical from.

Special Features
The developed system features some special characteristics that are not found in other information extraction systems. The system provides a module for automatic performance calculations with various predefined and well known parameters which are used to calculate performance of a system and also automatically draw a graph for graphical analysis of user. The systems also have a module for searching the results of extraction by any valid user in future if needed. The last but not least feature of the developed system is high precision ratio in each and every case of extraction this will became a major backbone of the developed IE system.

Implementation
At this stage the whole problems of unstructured data and extraction of useful information from large amount of data is well known. So concluding whole problems the concept of information extraction system seems to be a good idea where the useful information is easily extracted using some simple methods between all the three main methods of IE the rule based approach Is good idea. Our research focuses on the development of a reliable information extraction such as named entities such as a person, geographical names, designated post in work area etc. Rule based method of IE system mostly consist of a specification of a text basic rule language and interpreter which is able to apply the rules on documents in order to identify new information. In our algorithm first the text file is divided in the n-lines by the use of any strategy and then these lines will be divided into words chain such that we are able to differentiate each word is separated into lexical units. A lexical item is a single word, a part of a word, or a chain of words that forms the basic elements of a language's.

Top-down Approach:
In these approach firstly a generalized rule is defined which contain all the instances and cover all the instances and entity. In this starting rule is specialized to form new rules. Each specialization step ensures coverage of starting seed instances. This specialization is manual. In these coverage is very high but precision rate is low, and from these generalized rules specialized rules are created. And creation of rules is manual. We are using top-down approach and algorithm for these is:

a) Create rules of few instances.
b) Apply this on files(s).
c) Show Output.

In developed system, rules are created for each parameter and each pattern have a specific rule which is designed using the concepts of regular expressions and it can be applied to lexical units which are being created from the file. Rules can check on each and every individual string and matches each pattern and if the pattern matched with regard to any string then the string is relevant information for us that we try to extract.

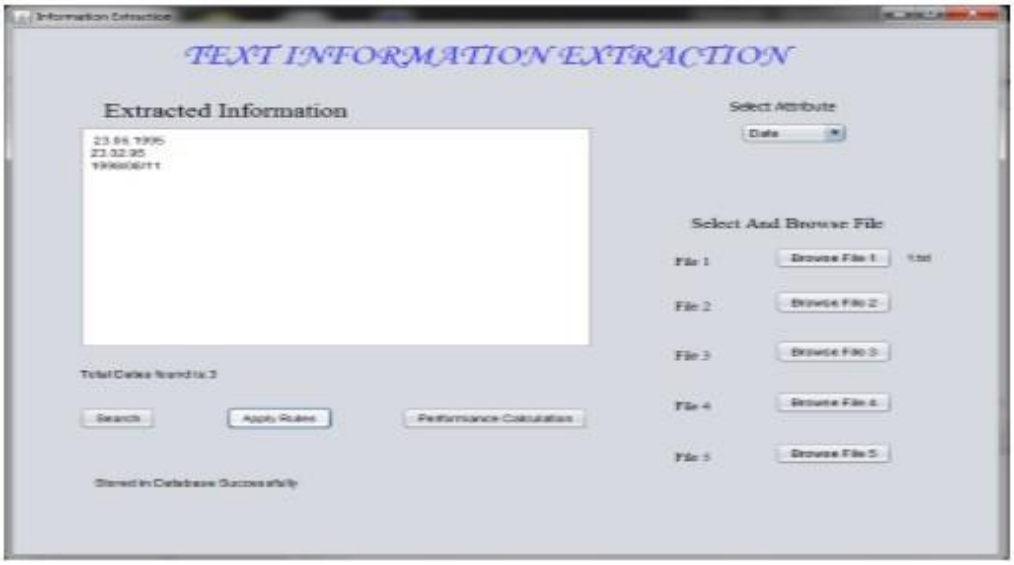

Figure 5.1: Implementation

Regular Expressions

A regular expression is a pattern that describes a set of strings. Regular expressions are constructed analogously to arithmetic expressions, by using various operators to combine smaller expressions. Some simple extraction tasks can be solved by writing regular expressions. Extraction from moderately more complex text sources, yet that has sufficient formatting regularity, can be addressed accurately with hand-tuned, programmed rules. For example, such rule "[a,p]m [0-9]+:[0-9]+" extract time description such as "AM 12:45" from documents. Regular expressions are widely used in Unix world as searching and replacing instruction. Using regular expressions for Information Extracting is similar as using regular expression for searching in Unix. Regular expression is suitable for such contents with significant syntactic properties (such as number, date etc.). In general regular expressions can be used for all document types as basic technique. Compared with other sophisticated methods, the processing of regular expressions is very quickly, because the only input of extracting rule is the defined regular expressions. No background knowledge or lexicon is required. Because regular expressions are based on Finite State Automata (FSA), learning and automatically generating regular expressions are theoretical possible. Although regular expressions locate fine information exactly, the contexts around the underlying fine information, which can help to locate information exactly, are not respected by using regular expressions alone. For instance, a regular expression "[0-9]+" defines all digit number sequence, but this rule is not able find out currency 100$ (which with suffix "$") exactly (assumed only the number has to be extracted but not the suffix). This shortcoming of regular expression causes sometimes very bad precision and recall when only standard regular expressions are used. Consequent of this shortcoming of regular expressions is that regular expressions have to be combined with other constraints, in order to ensure high performance IE. Hence, regular expressions are normally considered as basic technique and must be used with respect of context, in order to get a high performance extracting [14]. Pattern for name of person can be defined as -"{(Titles) ([Capitalized letter][small letters]{n}) ([Capitalized letter][small letters]{n}) }". It is applicable to such type of name "Mr. James Smith". And pattern for date can be given as "{[0-9](/)[0-9](/)[0-9]}". It is only applicable for one date pattern of "23/06/1995". There are n-number of formats of date and it can be defined manually of being generated then we defined more accurate common pattern as ([0-9]{2}(/|-|,|' ')[0-9]{2}(/|-|,|' ')[0-9]{4}) and many more as such patters are defined. Let take a example as "We both are research copartners namely Mr. Abhishek Patel and Mr. Ankit Patidar. We began our work on 02-March-2015 for more details contact, abc007@acropolis.in" So firstly these Text Can be divided in form of lines as 1st line - We both are research copartners namely Mr. Abhishek Patel and Mr. Ankit Patidar. 2nd line- We began our work on 02-March-2015 for more details contact, abc007@acropolis.in" And then these lines are divided into chain of string. Let take 1st line and it can be divided as "We| both| are| research| copartners| namely| Mr. |Abhishek| Patel| and| Mr.|Ankit| Patidar." "We| began| our| work| on| 02-March-2015| for| more| details| contact,| abc007@acropolis.in" now each string or word can be matched with the designed patterns and if matches then it is shown on output screen as extracted information.

Table 5.1 Resulting Table

So the extracted outcomes are:		
For 1st line Name: Date: Other Parameters:	 Mr. Abhishek Patel Mr. Ankit Patidar ---------- ----------	 (As names are matched with the pattern) (No date is there to match)
For 2nd line Name: Date: E-mail id: Other Parameters:	 --------- 02-March-2015 abc007@acropolis.in ----------	 (No name to match) (Date matched with pattern) (E-mail id matched with pattern)

RESULT ANALYSIS

In order to evaluate and analyze our system in more specific way, we use precision, recall and f-measure performance evolution criterions. In our experiment, the extraction of information is mainly based on 3 parameters namely name of a person, date, and e-mail address. The main reason for choosing that parameters is that all the chosen parameters exist in both our developed system which will referred as "IE system" and GATE also, so which will became the basis for comparison of our system to very famous system in field of text extraction. There are totally 8 files used for training process. All of them were randomly chosen from the internet and each file belongs to various fields. After completion of training process, the performance for each individual files is calculated separately for both the systems and average performance of each file on basis of various parameters are combined in following table

Table 6.1 Resulting figure

FILE NAME	Precision		Recall		F-measure	
	Gate	IE System	Gate	IE System	Gate	IE System
TEST 1.TXT	0.93	1	1	1	0.96	1
TEST 2.TXT	0.91	1	0.915	1	0.91	1
TEST 3.TXT	0.735	0.83	1	0.775	0.84	0.801
TEST 4.TXT	0.515	1	0.73	0.92	0.6	0.95
TEST 5.TXT	0.6	1	0.925	0.75	0.72	0.85
TEST 6.TXT	0.63	1	0.9	1	0.74	1
TEST 7.TXT	1	1	0.51	0.81	0.67	0.89
TEST 8.TXT	1	1	1	0.75	1	0.85

The graphical representation of various performance parameters are shown below which shows that the developed system is much more better than GATE system on comparison of precision measure and also in F-measure measurement to much extents, but it is somewhat lesser or equal to GATE in different test cases in sense of recall rate which is also considered as a main and important criterion.

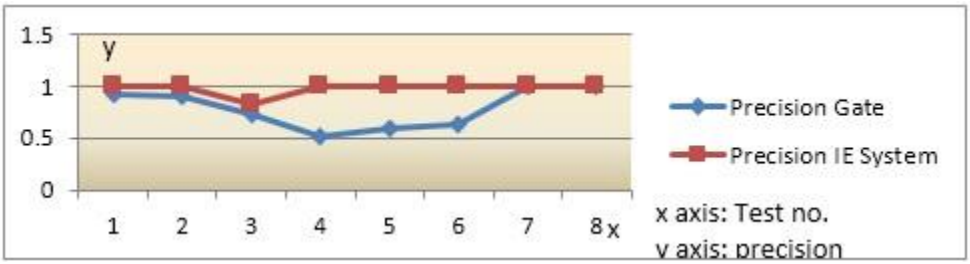

Figure 6.1Precision curve

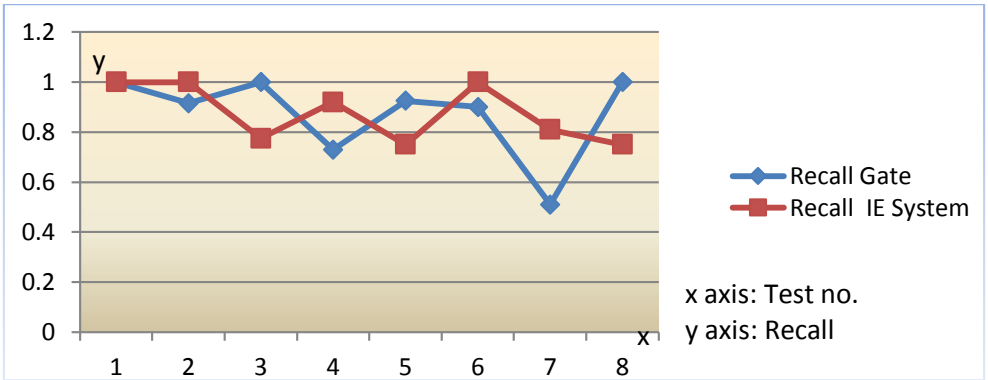

Figure 6.2 Recall curve

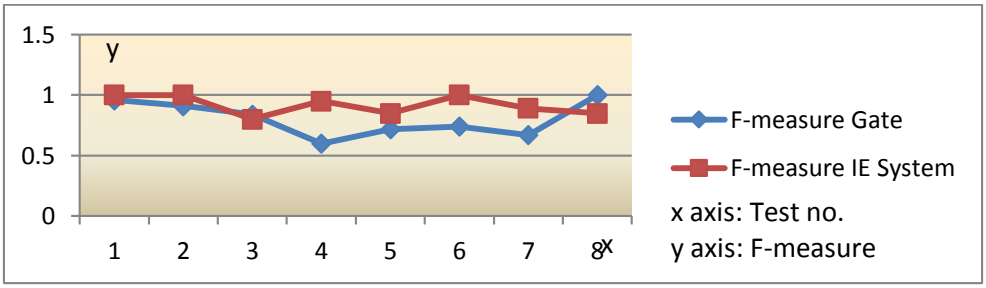

Figure 6.3 F-measure curve

The results of the experiment performed in the training process on the various training cases are combined and converges as follows if precision and F-measure is preferred, the developed system is good idea but with other parameter of recall the system performance is nearly same or little bit lesser than the existing system GATE.

CONCLUSION

In this research work we have discussed mainly one approach Rule Based information extraction for text documents. One can extract general named entity directly from text. As an example of this approach, we reviewed our project which extracted a information base of many different text file directly taken from internet or manually made. Second, we also calculate the various performance evolution criterions t judge the relevancy and accuracy in the results of extraction.

The main inference drawn from above research is that the developed system is much more better than its competitor in some extents but as every system have limitations too this developed system have limited parameters t extract for newer ones we have to make their rules also.

Research in information extraction continues to develop more effective algorithms for identifying entities and relations in text. By exploiting the latest techniques in human-language technology and computational linguistics and

combining them with the latest methods in top down approach f extraction, one can effectively extract useful and important information from the continually growing body both online and offline text files.

This research has presented a complete information extraction system not only for research projects but also for practical applications. Moreover, it is supported by a general methodology that allows construction of the IE system to be carried out in a focused and cost-effective way.

FUTURE WORK

Information extraction remains a challenging problem with many potential avenues for progress. In the developed system, we use a holistic approach for IE that addresses the limitations of the state-of the-art systems. This project concerns with the issues and efficient use of unsupported features in earlier systems developed, with an application to information extraction text files. This implemented IE system can itself be used to help improve extraction.
Our work can be extended in multiple directions. First, we consider using multi-objective techniques instead of relying on user-imposed constraints on recall and precision. Second, we plan to adapt our approach for modern parallel computation environments. In future this system also may be expanded to extract results from pdf files, obj. files, and other formats as well. We obtain new state-of-the-art performance in extracting standard fields from text files, with a significant performance calculation by several parameters. Developing semi-supervised learning methods for IE is a related research direction in which there has been only a limited amount of work. We also suggest better evaluation parameters to facilitate future research in this task. Especially in developing a system to extract information from images, videos, animations etc.

Our initial investigation shows that it is also promising for many applications. We are currently working on adapting IE techniques to the “Bottom up Approach”. In particular, we are working under the framework of a previously developed methodology to extract information from other file formats with more enhance feature. According to the common process model in information extraction, features are extracted from the input document and are used by a model to identify information. But using already extracted information for further information extraction can often account for missing or ambiguous features and increase the accuracy in domains with repetitive structure.

As another future work, more applications, especially practical applications, need to be investigated. The new applications can provide rich data sources for conducting information extraction, at the same time bring big challenges to the field. This is because various applications have various characteristics, needing to use different methods to deal with.

Fundamental advances in Rule Based Approach for Information extraction remain a significant open research area in future.

REFERENCES

[1] J. Tang, M. Hong, D. Zhang, B. Liang, and J. Li., “Information Extraction: Methodologies and Applications”. In the book of Emerging Technologies of Text Mining: Techniques and Applications, 2007
[2] S. Sarawagi,” Information Extraction”, Foundations and Trends R_ in Databases,Vol. 1, No. 3,2006
[3] R Mooney and R Bunescu,” Mining Knowledge from Text Using Information Extraction”, SIGKDD Explorations., Volume 7, Issue 1,2007
[4] C. Blaschke and A.Valencia.” The frame-based module of the Suiseki information extraction system.”, IEEE Intelligent Systems, 2002.
[5] Z. Zhang. ,”Weakly-Supervised Relation Classification for Information Extraction.” In Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM’2004), pp581-588, 2004.
[6] Ratnaparkhi, ,”Unsupervised Statistical Models for Prepositional Phrase Attachment.”, In Proceedings of COLING ACL’98. pp.1079-1085., 1998
[7] F. Sha,, & F. Pereira, ,”Shallow parsing with Conditional Random Fields.”, In Proceedings of Human Language Technology, NAACL. pp.188-191, 2003.
[8] T. Kristjansson, A.Culotta,, A. Viola , & McCallum ,”Interactive information extraction with constrained conditional random fields.” In Proceedings of AAAI’04, pp.412-418., 2004.
[9] Pinto, D., McCallum, A., Wei, X., & Croft, W. B. ,“Table Extraction Using Conditional Random Fields. “,In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’03).,pp. 235-242, 2003.

[10] Freitag, D. & McCallum, A. ,"Information Extraction with HMM Structures Learned by Stochastic Optimization.", In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI'2000), 2000.
[11] Muslea, I., Minton, S., & Knoblock, C. STALKER," Learning extraction rules for semistructured, web-based information sources.", In AAAI Workshop on AI and Information Integration. pp.74-81 , 1998
[12] Kushmerick, N., Weld, D. S., & Doorenbos, R. ,"Wrapper induction for information extraction.", In Proceedings of the International Joint Conference on Artificial Intelligence(IJCAI'97). pp.729-737, 1997
[13] Peter Kluegl, Martin Atzmueller, and Frank Puppe," TextMarker: A Tool for Rule-Based Information Extraction", 2009
[14] Ralph Grishman. Information Extraction: Techniques and Challenges Information Extraction (International Summer School SCIE-97), ed. Maria Teresa Pazienza, Springer-Verlag, 1997.
[15] D. M. Bikel, R. Schwartz, and R. M. Weischedel ,"An algorithm that learns what's in a name. Machine Learning",1999.
[16] Muawia Abdelmagid1, Ali Ahmed2 and Mubarak Himmat3," INFORMATION EXTRACTION METHODS AND EXTRACTION TECHNIQUES IN THE CHEMICAL DOCUMENT'S CONTENTS: SURVEY", ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 3, 2015
[17] MADINA IPALAKOVA,"INFORMATION EXTRACTION",2010