IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 16-20
www.iosrjournals.org

Information Extraction Technique: A Review

Varsha pande[1], Dr. A. S. Khandelwal[2]
[1](varshapande.var@gmail.com, Department of electronics and computer science, /RTMNU Nagpur, India)
[2](abha.ak@gmail.com,Department of computer science ,Hislop College ,/RTMNU Nagpur, India)

Abstract: *Now days, various digital information available electronically hasmade the organization of textual information into an important task. Text mining concerns looking for patterns in unstructuredtext. The Information Extraction (IE) is a technique that aims to extract the namesof entities and objects from text and to identify the rolesthat they play inevent descriptions. An IE system generally focusses on a specific domain or topic,searching only for information that is relevant to a user's interests.This paper presents the review of Information Extraction Technique.*

Keywords: *Text mining, Information extraction,*

I. INTRODUCTION

A huge amount of data available in digital form on the internet and intranets. A significant part of such information—e.g., government documents, corporate reports, online news, court rulings, legal acts, medical alerts and records, and social media communication—is transmitted through *unstructured*, free-text documents and is thus hard to search in. This resulted in a growing need for effective and efficient techniques for analyzing free-text data and discovering valuable and relevant knowledge from it in the form of *structured* information, and led to the emergence of Information Extraction technologies.

In the text mining chain, one of the initial links is information extraction. Its major goal is to transform the data from unstructured form into structured representation. The information extraction (IE) task can be expressed as to process the collection of texts which belong to a particular field and derive from each of them a previously defined set of name types, relations between them and events in which they participate. Each set of extracted entities is added, for instance, as a record to a table of a relational database in order that data mining techniques can be applied to this structured dataset later. There are two approaches to the information extraction system design, namely knowledge engineering and automatic training approaches. Both of them have their own benefits and drawbacks and are applied depending on the resources available to the system's designer.

II. TEXT MINING

Text mining [1] is defined as "the process of finding useful or interesting patterns, models, directions, trends, or rules from unstructured text". Several techniques have been proposed for text mining including information extraction, information retrieval, natural language processing, categorization, clustering, query processing etc. Information Retrieval (IR) techniques have been widely used for tasks such as document matching and ranking because a form of soft matching that utilizes word-frequency information typically gives superior results for most text processing problems

The architecture [2] of any text mining system containsthe following four main components:
1. Pre-processing which includes activities to prepare data to the next step. Typicallythey involve the process of converting the raw data from original source into theformat which is suitable for applying core mining operations.
2. Core mining operations which are the essence of the text mining technology. Theyprovide algorithms for pattern discovery in the data extracted from documents by thefirst component.The most widespread of them are distributions, frequent and nearfrequent sets and associations.
3. Presentation which provides a user interface with a query editor and visualization tools.
4. Refinement which includes optimization operations with the resulting data.

The pre-processing operations are divided into two broad categories which are techniquesaccording to their task and according to the algorithms and frameworks they employ. The firstgroup of approaches provides the structuring of the source documents and presenting them asthe task requires. The second group contains the approaches which imply the application offormal methods for analyzing available data. However, different techniques from bothcategories can be used in conjunction to solve many text mining tasks. Information

extractionis considered as a part of the task-oriented pre-processing approaches alongside withpreparatory processing and other natural language processing (NLP) techniques. While theother NLP and preparatory tasks can be defined as domain-independent, informationextraction itself is a highly domain-dependent technology.

Thereby, in the context of text mining technology information extraction can be classified asone of the pre-processing tasks which are used in order to make data ready for applying majordata mining techniques. These pre-processing operations involve processing the input,unstructured information in the form of documents, and presenting it in a more structured wayto make further post-processing analysis possible .

III. INFORMATION EXTRACTION

Even within the text mining [2]operations mention information extraction as the most important pre-processing technique which significantly increases the text mining potential. But moreover it is used as a self-dependent technology to settle the particular issues concerning the processing of the text information.

The information extraction [3]is used to find specific structured datain natural-language text. DARPA's Message Understanding Conferences(MUC) has concentrated on Information Extraction (IE) by evaluating the performance of participatingIE systems based on blind test sets of text documents .The data to be extracted is typically given by a template which specifies a list of slots to be filled with substrings taken from the document.Generally the data to be extracted is described by a template specifying a list of slots to be filled, though sometimes it is specified by annotations inthe document. Slot-fillers may be of two types: they may beeither one of a set of specified values or strings taken directly fromthe document.

Document
Title: Web Development Engineer
Location: Austin, TX
This individual is responsible for design and implementationof the web-interfacing components of the AccessBase server,and general back-end development duties.

A successful candidate should have experience that includes:
One or more of: Solaris, Linux, plus Windows/NT
Programming in C/C++, Java
Database access and integration: Oracle, ODBC
CGI and scripting: one or more of JavaScript,
Perl, PHP, ASP
Exposure to the following is a plus: JDBC, FrontPage and/orCold Fusion.

A BSCS and 2+ years experience (or equivalent) is required.
Filled Template

- title: \Web Development Engineer"
- location: \Austin, TX"
- languages: \C/C++", \Java", \JavaScript", \Perl", \PHP", \ASP"
- platforms: \Solaris", \Linux", \Windows/NT"
- applications: \Oracle", \ODBC", \JDBC", \FrontPage", \Cold Fusion"
- areas: \Database", \CGI", \scripting"
- degree required: \BSCS"
- years of experience: \2+ years"

Figure: Sample text and filled template for a job posting

The above Figure shows a shorteneddocument and template from aninformation extraction task in the job-posting domain. This template includesonly slots that are filled by strings taken directly from the document.Several slots may have multiple fillers for the job-posting domain asin (programming) languages, platforms, applications, and areas.IE has been shown to be useful in a variety of applications, e.g. restaurant guides, course homepages,seminar announcements, job postings, apartmentrental ads, and news articles on corporate acquisition. IE is also a suitable technology for automaticallyannotating web pages for the Semantic Web [4].

Machine learning techniques have been suggested for extractinginformation from text documents in order to create easily searchabledatabases from the information, thus making the online text more accessible [5]. For instance, information extracted from jobpostings on the Web can be used to build a searchable database of jobs.

IV. OVERVIEW OF IE BASED TEXT MINING FRAMEWORK

Traditional data mining assumes that the information to be "mined" is already in the form of arelational database. Unfortunately, for many applications, electronic information is only availablein the form of free natural-language documents rather than structured databases. Since IE addressesthe problem of transforming a corpus of textual documents into a more structured database, thedatabase constructed by an IE module can be provided to the KDD module for further mining ofknowledge as illustrated in Figure 1. Information extraction can play an obvious role in text miningas illustrated.

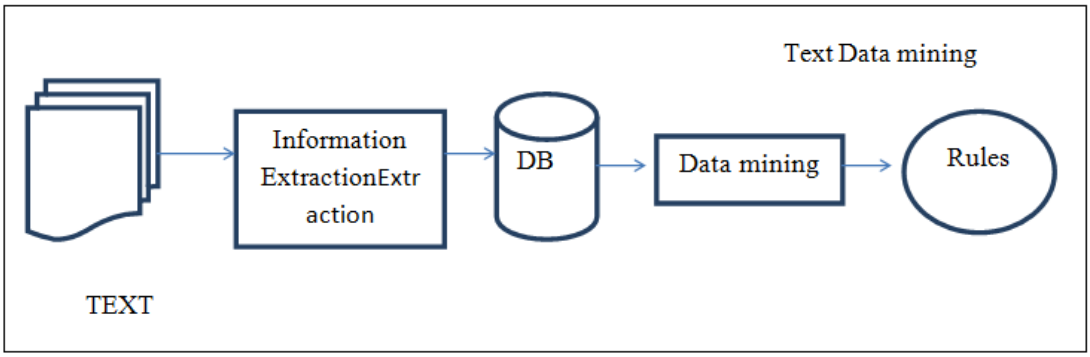

Figure 1: Overview of IE-based text mining framework

Although constructing an IE system is a difficult task, there has been significant recent progressin using machine learning methods to help automate the construction of IE systems.

By manually annotating a small number of documents with the information to be extracted, areasonably accurate IE system can be induced from this labeled corpus and then applied to a largecorpus of text to construct a database. However, the accuracy of current IE systems is limitedand therefore an automatically extracted database will inevitably contain significant numbers oferrors. An important question is whether the knowledge discovered from this "noisy" database issignificantly less reliable than knowledge discovered from a cleaner database. This paper presentsexperiments showing that rules discovered from an automatically extracted database are close inaccuracy to that discovered from a manually constructed database.

V. THE OVERALL PROCESS OF INFORMATION EXTRACTION

Different authors divide the process of information extraction in different steps of different granularity, combining them into bigger stages and assigning the components of the information extraction systems to accomplish the tasks involved[5]. However, analyzing those different approaches the general pipeline of the information extraction process [2] can be summarized. In the current work six main stages were determined as following:

1. Initial processing.
2. Proper names identification.
3. Parsing.
4. Extraction of events and relations.
5. Anaphora resolution.
6. Output results generation.

1 Initial processing

There are several operations which usually compose the primary step of the information extraction process. The first of them is the splitting a text into the fragments which are defined differently throughout the papers from different researchers like zones, sentences, segments or tokens. This procedure can be performed by the components named as tokenisers, text zoners, segmenters or splitters. A tokenization [7] is a quite straightforward task for the texts in any European language, where the blank space between characters and punctuation indicate the boundaries of a word and a sentence respectively. But, for example, for Chinese or Japanese texts, where the boundaries are not so obvious this operation is not the simple one and requires much more effort to fulfill it.

The next task within the initial processing stage is usually the morphological analysis whichincludes part-of-speech tagging and phrasal units (noun or verb phrases) identification. Partof-speech tagging might be helpful to the next step which is the lexical analysis. It handlesunknown words and resolves ambiguities, some of them by identifying part-of-speech of thewords which cause those ambiguities. In addition, the lexical analysis involves working withthe specialised dictionaries and gazetteers, which are composed of different types

of names:titles, countries, cities, companies and their suffixes, positions in a company, etc. If a word ina document is found in a gazetteer it is tagged with the semantic class the word belongs to.

For example, a word "Mr" will be tagged with the semantic class "Titles".

Some authors add a filtering task to the pre-processing stage which implies selecting onlythose sentences which are relevant to the extraction requirements [8].

2 Proper names identification

One of the most important operations in the chain ofinformation extraction is the identification of various classes of proper names, such as namesof people or organisations, dates, currency amounts, locations, addresses, etc. They can beencountered in almost all types of texts and usually they constitute the part of the extractionscenario. These names are recognised using a number of patterns which are called regularexpressions[9]. However, usually authors do not classify thisoperation as a separate task within the whole information extraction process.

3 Parsing

During this stage the syntactic analysis of the sentences in the documents isperformed. After the previous step, where the basic entities were recognised the sentences areparsed to identify the noun group around some of those entities and verb groups. This parsingstage must be done in order to prepare the ground for the next stage of extraction relationsbetween those entities and events in which they participate. The noun and verb groups are used as sections to begin to work on at the pattern matching stage. The identification of thosegroups is realised by applying a set of specially constructed regular expressions [6].

However, the full parsing is not an easy task; therefore it requires expensive computations tobe involved which in its turn slow down the whole process of information extraction. Since itis a difficult problem, the full parsing is prone to introduce errors. In contrast, sometimes thefull syntactic analysis might not be needed at all. Thereby, more and more informationextraction research groups tend to use so called partial or shallow parsing instead of full one.Using only local information the shallow parsing creates partial, not overlapping syntacticfragments which are identified with a higher level of confidence. At the beginning of theevaluation process all of the MUC's participants used the full parsing. And the group thatcame up with the new idea of shallow parsing was Lehnert*et. al*. during MUC-3 in 1991. Asa result of applying the partial syntactic analysis, they showed a better performance than therest of the sites which tried to create full syntactic structures [6].

4 Extraction of events and relations

Everything which is done previously is basically the preparation for the major stage of extraction of events and relations, which are particularlyrelated to the initial extraction specifications given by a client. This process is realised bycreating and applying extraction rules which specify different patterns. The text is matchedagainst those patterns and if a match is found the element of the text is labelled and laterextracted. The formalism of writing those extraction rules differs from one informationextraction system to another[7].

5 Anaphora resolution

Despite the fact that this problem was firstly introduced and evaluatedon the MUC-6 as the coreference task (CO), before the MUC-6 coreference presented as achallenge and research groups tried to resolve it, although implicitly.Any given entity in a text can be referred to several times and every time it might be referreddifferently. In order to identify all the ways used to name that entity throughout the documentcoreference resolution is performed. Coreference or anaphora resolution is the stage when fornoun phrases it is determined if they refer to the same entity or not. There are several types ofcoreference, but the most common types are pronominal and proper names coreference, whena noun is replaced by a pronoun in the first case and by another noun or a noun phrase in thesecond one[9].

6 Output results generation

This stage involves transforming the structures which wereextracted during the previous operations into the output templates according to the formatspecified by a client. It might include different normalisation operations for dates, time,currencies, etc. For instance, a round-off procedure for percentages can be executed and areal number 75.96 will be turned into integer 76[8].

Not all of the tasks must me necessarily accomplished within one information extractionproject. Therefore, a particular information extraction system does not have to have all ofthose possible components. According to Appelt and Israel there are several factorsthat affect the choice of systems' components, like:

- **Language** : As it was mentioned earlier for processing texts in Chinese or Japaneselanguages with not clear word and sentence boundaries or texts in German languagewith words of a difficult morphological structure some modules are definitelynecessary compared to working with English documents.
- **Text genre and properties**: In transcripts of informal speech, for example, spellingmistakes might occur in addition to implicit sentence boundaries. If information mustbe extracted from such texts those issues must be taken into consideration andaddressed while designing a system by adding corresponding modules.
- **Extraction task**: For an easy task like names recognition the parsing and anaphoraresolution modules might not be needed at all.

VI. EVALUATION IN INFORMATION EXTRACTION

Given an input text, or a collection of texts, the expected output of an IE system canbe defined very precisely. This facilitates the evaluation [10] of different IE systems andapproaches. In particular, the *precision* and *recall* metrics were adopted from theIR research community for that purpose. They measure the system's effectivenessfrom the user's perspective, i.e., the extent to which the system produces allthe appropriate output (recall) and only the appropriate output (precision). Thus,recall and precision can be seen as measure of completeness and correctness,respectively. To define them formally, let *#key* denote the total number of slotsexpected to be filled according an *annotated* reference corpus, representing groundtruth or a "gold-standard", and let *#correct* (*#incorrect*) be the number of correctly(incorrectly) filled slots in the system's response. A slot is said to be filledincorrectly either if it does not align with a slot in the gold standard (*spurious slot*)or if it has been assigned an invalid value. Then, precision and recall may be definedas follows:

$$\textbf{Precision}=\frac{\#correct}{\#correct+\#incorrect} \qquad \textbf{recall}=\frac{\#correct}{\#key} \qquad (2.1)$$

In order to obtain a more fine-grained picture of the performance of IE systems,precision and recall are often measured for each slot type separately.
The *f-measure* is used as a weighted harmonic mean of precision and recall,which is defined as follows:

$$\textbf{F}=\frac{(\beta^2+1)\times\textbf{precision}\times\textbf{recall}}{(\beta^2\times precision)+recall} \qquad (2.2)$$

In the above definition β is a non-negative value, used to adjust their relativeweighting (β^2=1.0 gives equal weighting to recall and precision, and lower valuesof βgive increasing weight to precision).
Other metrics are used in the literature as well, e.g., the so called *slot error rate*,SER [40], which is defined as follows:

$$\textbf{SER}=\frac{\#incorrect+\#missing}{\#key} \qquad (2.3)$$

where *#missing* denotes the number of slots in the reference that do not align withany slots in the system response. It reflects the ratio between the total number of sloterrors and the total number of slots in the reference. Depending on the particularneeds, certain error types, (e.g., spurious slots) may be weighted in order to deemthem more or less important than others.

VII. CONCLUSION

Information Extraction seems to be the best technique for extracting the text using text mining. The overall process of information extraction based on text mining framework is studied. The factors such as precision, Recall, F-measure and slot error rate are used for evaluation in information extraction are also discussed.

REFERENCES:

[1]. Mooney, U. Y. N. a. R. J., 2002. Text Mining with Information Extraction, Austin: s.n.
[2]. Ipalakova, M., 2010. INFORMATION EXTRACTION, s.l.: UNIVERSITY OF MANCHESTER.
[3]. Nahm, U. Y., n.d. Text Mining with Information Extraction. s.l.:s.n.
[4]. Berners-Lee T, H. J. &. L. O., 2001. The Semantic Web ,Scientific american. s.l., s.n.
[5]. califf M. E., a. M. R. J., 1999. Relational learning of pattern-match rules for information extraction. Orlando, FL, s.n.
[6]. Grishman, R., 1997. Information Extraction: Techniques and Challenges. Berlin, Heidelberg, Springer-Verlag, pp. 10-27.
[7]. D, A. D. a. I., 1999. Introduction to Information Extraction Technology. IJCAI-99.
[8]. J Turmo, A. A. a. N. C., 2006. Adaptive Information Extraction. ACM computing surveys, pp. 1-47.
[9]. Feldman, R. .., 2007. The Text Mining Handbook: Advanced Approaches In Analyzing Unstructured Data. New York, Cambridge University Press.
[10]. Yangarber, J. P. a. R., 2013. Information Extraction: Past, Present and future. In: Theory and Applications of Natural Language Processing. Verlag Berlin Heidelberg: Springer, pp. 23-48.