Information Extraction Technique: A Review

Varsha pande¹, Dr. A. S. Khandelwal²

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 has made the organization of textual information into an important task. Text mining concerns looking for patterns in unstructured text. The Information Extraction (IE) is a technique that aims to extract the names of entities and objects from text and to identify the roles they play invent descriptions. An IE system generally focuses 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 contains the following four main components:
1. Pre-processing which includes activities to prepare data to the next step. Typically they involve the process of converting the raw data from original source into the format which is suitable for applying core mining operations.
2. Core mining operations which are the essence of the text mining technology. They provide algorithms for pattern discovery in the data extracted from documents by the first component. The most widespread of them are distributions, frequent and near-frequent 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 techniques according to their task and according to the algorithms and frameworks they employ. The first group of approaches provides the structuring of the source documents and presenting them as the task requires. The second group contains the approaches which imply the application of formal methods for analyzing available data. However, different techniques from both categories can be used in conjunction to solve many text mining tasks. Information
Information Extraction Technique: A Review

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

Thereby, in the context of text mining technology information extraction can be classified as one of the pre-processing tasks which are used in order to make data ready for applying major data mining techniques. These pre-processing operations involve processing the input, unstructured information in the form of documents, and presenting it in a more structured way to 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 data in natural-language text. DARPA's Message Understanding Conferences (MUC) has concentrated on Information Extraction (IE) by evaluating the performance of participating IE 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 in the document. Slot-fillers may be of two types: they may be either one of a set of specified values or strings taken directly from the document.

Document
Title: Web Development Engineer
Location: Austin, TX
This individual is responsible for design and implementation of 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
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 shortened document and template from an information extraction task in the job-posting domain. This template includes only slots that are filled by strings taken directly from the document. Several slots may have multiple fillers for the job-posting domain as in (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, apartment rental ads, and news articles on corporate acquisition. IE is also a suitable technology for automatically annotating web pages for the Semantic Web [4].

Machine learning techniques have been suggested for extracting information from text documents in order to create easily searchable databases from the information, thus making the online text more accessible [5]. For instance, information extracted from job postings 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 a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural-language documents rather than structured databases. Since IE addresses the problem of transforming a corpus of textual documents into a more structured database, the database constructed by an IE module can be provided to the KDD module for further mining of knowledge as illustrated in Figure 1. Information extraction can play an obvious role in text mining as illustrated.

![Figure 1: Overview of IE-based text mining framework](image)

Although constructing an IE system is a difficult task, there has been significant recent progress in 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, a reasonably accurate IE system can be induced from this labeled corpus and then applied to a large corpus of text to construct a database. However, the accuracy of current IE systems is limited and therefore an automatically extracted database will inevitably contain significant numbers of errors. An important question is whether the knowledge discovered from this “noisy” database is significantly less reliable than knowledge discovered from a cleaner database. This paper presents experiments showing that rules discovered from an automatically extracted database are close in accuracy 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 zone rs, 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 which includes part-of-speech tagging and phrasal units (noun or verb phrases) identification. Part-of-speech tagging might be helpful to the next step which is the lexical analysis. It handles unknown words and resolves ambiguities, some of them by identifying part-of-speech of the words which cause those ambiguities. In addition, the lexical analysis involves working with the 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 in a 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 only those sentences which are relevant to the extraction requirements [8].

2 Proper names identification

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

3 Parsing

During this stage the syntactic analysis of the sentences in the documents is performed. After the previous step, where the basic entities were recognised the sentences are reparsed to identify the noun group around some of those entities and verb groups. This parsing stage must be done in order to prepare the ground for the next stage of extraction relations between those entities and events in which they participate. The noun and verb groups are used as sections to begin work on at the pattern matching stage. The identification of those groups 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 to be involved which in its turn slow down the whole process of information extraction. Since it is a difficult problem, the full parsing is prone to introduce errors. In contrast, sometimes the full syntactic analysis might not be needed at all. Thereby, more and more information extraction 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 syntactic fragments which are identified with a higher level of confidence. At the beginning of the evaluation process all of the MUC’s participants used the full parsing. And the group that came up with the new idea of shallow parsing was Lehnert et al. during MUC-3 in 1991. As a result of applying the partial syntactic analysis, they showed a better performance than the rest 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 particularly related to the initial extraction specifications given by a client. This process is realised by creating and applying extraction rules which specify different patterns. The text is matched against those patterns and if a match is found the element of the text is labelled and later extracted. The formalism of writing those extraction rules differs from one information extraction system to another [7].

5 Anaphora resolution

Despite the fact that this problem was firstly introduced and evaluated on the MUC-6 as the coreference task (CO), before the MUC-6 coreference presented as a challenge 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 referred differently. In order to identify all the ways used to name that entity throughout the document coreference resolution is performed. Coreference or anaphora resolution is the stage when pronoun phrases it is determined if they refer to the same entity or not. There are several types of coreference, but the most common types are pronominal and proper names coreference, when a noun is replaced by a pronoun in the first case and by another noun or a noun phrase in the second one [9].

6 Output results generation

This stage involves transforming the structures which were extracted during the previous operations into the output templates according to the format specified 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 a real number 75.96 will be turned into integer 76 [8].

Not all of the tasks must be necessarily accomplished within one information extraction project. Therefore, a particular information extraction system does not have to have all of those possible components. According to Appelt and Israel there are several factors that affect the choice of systems’ components, like:
- **Language**: As it was mentioned earlier for processing texts in Chinese or Japanese languages with not clear word and sentence boundaries or texts in German language with words of a difficult morphological structure, some modules are definitely necessary compared to working with English documents.

- **Text genre and properties**: In transcripts of informal speech, for example, spelling mistakes might occur in addition to implicit sentence boundaries. If information must be extracted from such texts those issues must be taken into consideration and addressed while designing a system by adding corresponding modules.

- **Extraction task**: For an easy task like names recognition the parsing and anaphora resolution 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 can be defined very precisely. This facilitates the evaluation [10] of different IE systems and approaches. In particular, the precision and recall metrics were adopted from the IR research community for that purpose. They measure the system’s effectiveness from the user’s perspective, i.e., the extent to which the system produces all the 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 # keys denote the total number of slots expected to be filled according an annotated reference corpus, representing ground truth 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 filled incorrectly 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 defined as follows:

\[
\text{Precision} = \frac{\#\text{correct}}{\#\text{correct} + \#\text{incorrect}} \quad \text{recall} = \frac{\#\text{correct}}{\#\text{key}}
\]

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:

\[
F_\beta = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}}
\]

In the above definition, $\beta$ is a non-negative value, used to adjust their relative weighting ($\beta^2 = 1.0$ gives equal weighting to recall and precision, and lower values of $\beta$ give 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:

\[
\text{SER} = \frac{\#\text{incorrect} + \#\text{missing}}{\#\text{key}}
\]

where $\#\text{missing}$ denotes the number of slots in the reference that do not align with any slots in the system response. It reflects the ratio between the total number of slot errors and the total number of slots in the reference. Depending on the particular needs, certain error types, (e.g., spurious slots) may be weighted in order to deem them 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: