An Overview of Approaches to Extract Information from Natural Language Corpora

Frederik Hogenboom
fhogenboom@ese.eur.nl

Flavius Frasincar
frasincar@ese.eur.nl

Uzay Kaymak
u.kaymak@ieee.org

Econometric Institute
Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands

ABSTRACT

It becomes increasingly important to be able to handle large amounts of data more efficiently, as anyone could need or generate a lot of information at any given time. However, distinguishing between relevant and non-relevant information quickly, as well as responding to newly obtained data of interest adequately, remain cumbersome tasks. Therefore, a lot of research aiming to alleviate and support the increasing need of information by means of Natural Language Processing (NLP) has been conducted during the last decades. This paper reviews the state-of-the-art of approaches on information extraction from text. A distinction is made between statistic-based approaches, pattern-based approaches, and hybrid approaches to NLP. It is concluded that it depends on the user's need which method suits best, as each approach to natural language processing has its own advantages and disadvantages.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*knowledge acquisition*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*; I.5.4 [**Pattern Recognition**]: Applications—*text processing*

General Terms

Languages, algorithms

Keywords

Information extraction, natural language processing (NLP), text mining, parsing

1. INTRODUCTION

In today's busy, data-driven world where the stream of vital information never ends, anyone could generate or collect a lot of information about almost anything at any given time. Due to the enormous and ever growing amount of information that is available, it becomes increasingly important to be able to handle these large amounts of data more efficiently. For instance, one has to be able to distinguish between relevant and non-relevant information quickly and respond to newly obtained data of interest adequately.

As a consequence, a lot of research aiming to support the increasing need of information has been conducted during the last decades. At first, research was mainly focused on Information Retrieval (IR), but currently the focus has shifted to Information Extraction (IE) or data mining. An omnipresent problem is the fact that most data is unstructured, being described using natural (human-understandable) language. In order to retrieve or process large amounts of data, it is desired to make use of machines. However, natural languages are not machine-understandable and thus there is a need for performing Natural Language Processing (NLP) [2].

NLP is a field in computer science and linguistics that is closely related to Artificial Intelligence (AI) and Computational Linguistics (CL). NLP is generally employed to convert information stored in natural language to a machine-understandable format. Thus, the main goal of NLP and IE is to extract knowledge from unstructured data. The main difficulties that are encountered with NLP arise when longer sentences that are highly ambiguous and have complex grammars are to be processed. The challenges imposed by automatically processing natural language have motivated the ongoing research into NLP for several decades.

Throughout the years, many NLP systems have been created, and nowadays, innovative NLP systems are still being developed, as the popularity of NLP witnesses a substantial growth, caused by, for example, the huge amount of available (electronic) text and the presence of adequate processing power. NLP systems vary in employed techniques, are built for different purposes, and may differ in focus. In general, one can distinguish between several layers within the processing tasks performed by an NLP system [1]. These levels of language range from the phonology and morphology of elements, to the lexical, syntactic, and semantic aspects of text, to the discourse and pragmatic properties of natural language text. Some systems focus more on the lower levels of processing, whereas other systems focus on the higher levels or on all levels. Generally speaking, three main approaches to NLP exist, i.e., statistics-based, pattern-based, and hybrid approaches.

DIR 2010 January 25, 2010, Nijmegen, the Netherlands.

2. STATISTICS-BASED APPROACHES

Statistical approaches are commonly used for natural language processing applications. These methods are data-driven and rely solely on (automated) quantitative methods to discover relations. Statistical approaches require large text corpora in order to develop models that approximate linguistic phenomena. Furthermore, statistics-based NLP is not restricted to basic statistical reasoning based on probability theory, but encompasses all quantitative approaches to automated language processing, such as probabilistic modeling, information theory, and linear algebra.

Even though one could distinguish between word-based and grammar-based approaches (e.g., word frequency counting and part-of-speech tagging, respectively), all statistics-based approaches to NLP share their focus on discovering statistical relations, i.e., facts that are supported by statistical evidence. Examples of discovered facts are words or concepts that are (statistically) associated with one another. It should be noted that statistical relations do not necessarily imply semantically valid relations or relations that have proper semantic meaning.

Hence, statistical methods do not deal with meaning explicitly, i.e., they discover relations in corpora without considering semantics. However, from a statistical point of view, this is a matter of definition more than it is a real issue. One could argue that true meaning is not related to philosophical semantics, but to evidence that resides within the distribution of contexts over which words and utterances are used. Another disadvantage of statistics-based NLP is that it requires a large amount of data in order to result in statistically significant results. However, these approaches are not based on knowledge, and thus neither linguistic resources, nor expert knowledge are required.

3. PATTERN-BASED APPROACHES

In contrast to statistics-based approaches, pattern-based approaches are based on linguistic or lexicographic knowledge, as well as existing human knowledge regarding the contents of the text that is to be processed. This knowledge is mined from corpora by using predefined or discovered patterns. One could distinguish different types of patterns, i.e., lexico-syntactic and lexico-semantic patterns. The former patterns combine lexical representations and syntactical information with regular expressions, whereas the latter patterns also employ semantic information. These semantics are added by means of gazetteers (which use the linguistic meaning of text) or ontologies (which also include relationships).

There are several advantages that result from the utilization of pattern-based approaches to perform NLP tasks over statistics-based approaches. First of all, pattern-based approaches need less training data than statistical NLP approaches. Also, it is possible to define powerful expressions by using lexical, syntactical, and semantic elements, and results are easily interpretable. Patterns are useful when one needs to extract very specific information. However, in order to be able to define patterns that retrieve the correct, desired information, lexical knowledge and possibly also prior domain knowledge is required. Other disadvantages are related to defining and maintaining patterns, as these are cumbersome and non-trivial tasks.

4. HYBRID APPROACHES

Although theoretically there is a crisp distinction between statistical and pattern-based approaches, in reality, it appears to be difficult to stay within the boundaries of a single approach. Often, an approach to NLP can be considered as mainly statistical or pattern-based, but there is also an increasing number of researchers that equally combine data-driven and knowledge-driven approaches, to which we refer to as hybrid approaches. For instance, it is hard to apply solely pattern-based algorithms successfully, as these algorithms often need for instance bootstrapping or initial clustering, which can be done by means of statistics. Furthermore, hybrid approaches to NLP could emerge when solving the lack of expert knowledge for pattern-based approaches, by applying statistical methods. Also, researchers can combine statistical approaches with (lexical) knowledge, for instance to prevent unwanted results.

By combining different techniques, advantages as well as disadvantages of statistical and pattern-based approaches are inherited. For instance, one is able to create complex patterns and less data is required compared to statistical approaches, but more data is required compared to pattern-based methods. Some inherited disadvantages can be (partially) cleared by advantages, e.g., the lack of semantics in statistical methods is solved when adding patterns. Disadvantages of hybrid approaches to NLP are related to the multidisciplinary aspects of hybrid NLP systems.

5. CONCLUSION

In this survey, we have elaborated on the main approaches to natural language processing. Each of the approaches has its advantages and disadvantages, and thus we can define guidelines regarding the selection of a proper NLP approach. If one is less concerned with semantics and assumes that knowledge lies within statistical facts on a specific corpus, it is advised to use statistics-based approach. Otherwise, if one is concerned with the semantics of discovered information, or it is desired to be able to easily explain and control the results, a pattern-based approach is more suitable. However, if one needs to bootstrap a pattern-based approach using statistics (for instance when there is insufficient expert knowledge available) or the other way around (e.g., when there is a need for a priori knowledge), a hybrid approach is more appropriate.

6. REFERENCES

[1] E. D. Liddy. *Encyclopedia of Library and Information Science*, chapter Natural Language Processing, pages 2126–2136. Marcel Decker, Inc., 2nd edition, 2003.

[2] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1st edition, 1999.