Erasmus University Rotterdam

An Overview of Approaches to Extract Information from Natural Language Corpora

Natural Language Processing

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.

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. Generally speaking, three main approaches to NLP exist, i.e., **statistics-based**, **pattern-based**, and **hybrid** approaches.

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 statistical 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 (**word-** and **grammar-based**) quantitative approaches to automated language processing, such as probabilistic modeling, information theory, and linear algebra.

Advantages of these approaches are:

- neither linguistic resources, nor expert knowledge are required;
- issues regarding leaking grammars, inconsistencies among humans, dialects, etc. are alleviated.

Disadvantages of these approaches are:

- often a substantial amount of data is needed;
- these approaches do not deal with meaning (semantics) explicitly.

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 between several patterns, i.e., **lexico-syntactic** and **lexico-semantic patterns**. Lexico-syntactic 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 the text) or ontologies (which also include relationships).

Advantages of these approaches are:

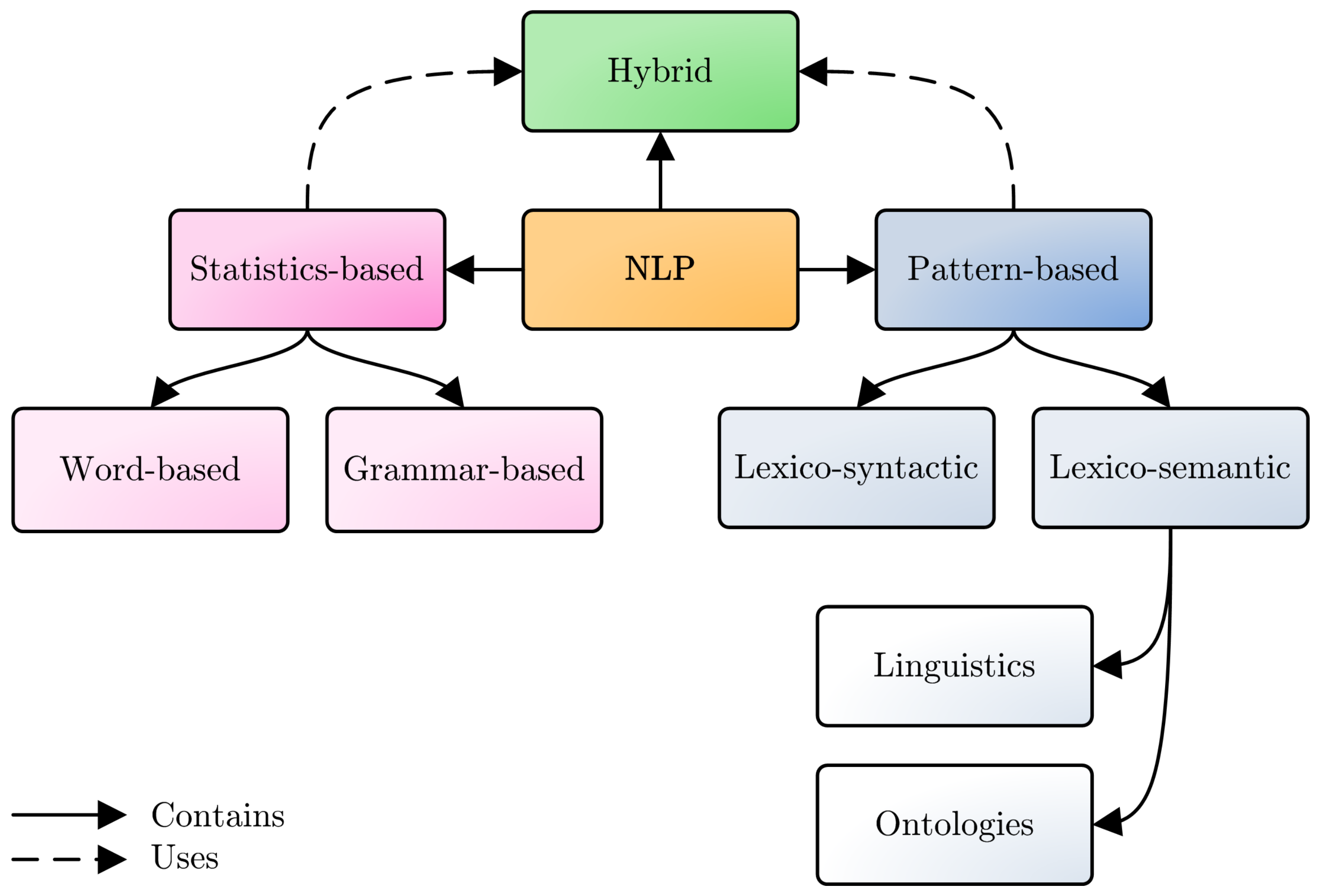
- less training data is needed;
- one can define powerful expressions;
- results are easily interpretable.

Disadvantages of these approaches are:

- the requirement of lexical and possibly also domain knowledge;
- defining and maintaining patterns are often cumbersome and non-trivial tasks.

Frederik Hogenboom
fhogenboom@ese.eur.nl
Flavius Frasincar
frasincar@ese.eur.nl
Uzay Kaymak
u.kaymak@ieee.org

Econometric Institute
Erasmus School of Economics
Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands
http://www.eur.nl/english/



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. Also, researchers can **combine statistical** approaches with **lexical knowledge**. Furthermore, hybrid approaches to NLP could emerge when **solving the lack of expert knowledge problem** for pattern-based approaches, by applying statistical methods.

Advantages of these approaches are:

- problems related to scaling and required expert knowledge of pattern-based approaches are addressed;
- not as much data as needed for statistical approaches is required;
- semantics are dealt with.

Disadvantages of these approaches are:

- due to the combination of techniques, maintaining completeness and accuracy of the system becomes more difficult;
- multidisciplinary aspects require special care.

Conclusions

As each of the approaches has its advantages and disadvantages, **guidelines** regarding the selection of a proper NLP approach can be defined:

- if semantics are not a concern and it is assumed that knowledge lies within statistical facts on a specific corpus, a statistics-based approach should be used;
- if the semantics of discovered information are a concern, or it is desired to be able to easily explain and control the results, a pattern-based approach is suitable;
- if bootstrapping a pattern-based approach using statistics (e.g., insufficient knowledge available) is needed, or the other way around (e.g., need of a priori knowledge), a hybrid approach should be considered.