Text-Based Information Extraction Using Lexico-Semantic Patterns

Frederik Hogenboom Wouter IJntema Flavius Frasincar

Erasmus University Rotterdam
P.O. Box 1738, NL-3000 DR, Rotterdam, the Netherlands
fhogenboom@ese.eur.nl, wouterijntema@gmail.com, frasincar@ese.eur.nl

The full version of this paper, entitled A Lexico-Semantic Pattern Language for Learning Ontology Instances from Text, will appear in Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Elsevier, 2012 (DOI: 10.1016/j.websem.2012.01.002)

Abstract

In order to cope with the ever increasing amount of data available on the Web, information extraction patterns are frequently employed to gather relevant information. Currently, most patterns use lexical and syntactic elements, but fail to exploit domain semantics. We propose a lexico-semantic pattern-based rule language, i.e., the Hermes Information Extraction Language (HIEL), which exploits a domain ontology for pattern creation. Experiments on financial news show that HIEL rules outperform lexico-syntactic rules and state-of-the-art lexico-semantic JAPE rules in terms of rule creation times and F_1 scores.

1 Introduction

The tremendous growth of the Web has resulted in enormous amounts of data that are readily available to the average user. Many researchers have hence developed ways to convert these vast amounts of data into valuable information that can be used for various purposes, e.g., decision support or trading tools. For information extraction, patterns are frequently applied. For example, simple lexico-syntactic patterns [3] can be used to extract hyponyms from text. The problem with these type of patterns is that their support is often limited to hypernym, hyponym, meronym, and holonym relations. They employ limited syntactical elements and do not make use of the domain semantics. While these patterns generate high precision, recall lags behind, hence driving the development of lexico-semantic pattern languages like JAPE [1] to cope with this issue. However, these often suffer from verbosity and complexity, or semantic elements are not exploited to their full potential (e.g., by making use of a reasoning engine).

Therefore, we introduce the lexico-semantic pattern-based rule language called the Hermes Information Extraction Language (HIEL), which makes use of lexical and syntactical elements, as well as semantic elements. HIEL utilizes Semantic Web technologies by employing domain ontologies, hereby exploiting the domain concepts through the use of inference. Our language is evaluated in the Hermes news processing framework [2], of which the underlying ontology consists of lexicalized concepts for the financial domain.

2 Language Syntax

Figure 1 shows an example rule that links CEOs to their subjective companies, to illustrate the main features of our language. Lexical and syntactic elements are indicated by white labels, whereas semantic elements (which make use of the Hermes knowledge base) are indicated by shaded labels.

In HIEL, a rule typically consists of a left-hand side (LHS) and a right-hand side (RHS). Once the pattern on the RHS has been matched, it is used in the LHS, which consists of three components, i.e., a subject, predicate, and an object, where the predicate describes the relation between the subject and the object (in this case `hasCEO`). The RHS supports sequences of many different features, as explained below.

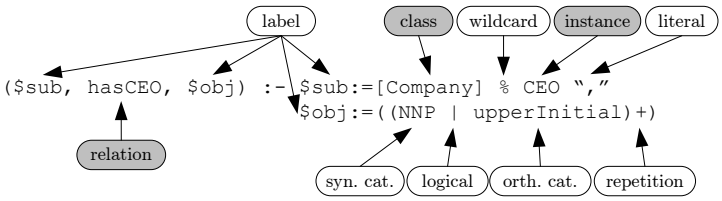

Figure 1: Example HIEL pattern

First, labels (preceded by `$`) on the RHS associate sequences using `:=` to the correct entities specified on the LHS. Second, syntactic categories (e.g., nouns, verbs, etc.) and orthographical categories (i.e., token capitalization) can be employed. Next, HIEL supports the basic logical operators *and* (`&`), *or* (`|`), and *not* (`!`), and additionally allows for repetition (regular expression operators, i.e., `*`, `+`, `?`, and $\{\ldots\}$). Moreover, wildcards are also supported, allowing for ≥ 0 tokens (`%`) or exactly 1 token (`_`) to be skipped.

Of paramount importance is the support for semantic elements through the use of concepts, i.e., ontological classes, which are defined as groups of individuals that share the same properties, i.e., the instances of a class. A concept may consist of alternative lexical representations that are stored using the synonym property. The hierarchical structure of the ontology allows the user to make rules either more specific or more generic, depending on the needs at hand.

3 Evaluation

We have implemented our language as an extension to the Hermes News Portal (HNP) [2] by adding a HIEL rule engine, rule editor, and annotation validator. In our experiments we use 500 financial news articles scraped from the Web and an ontology consisting of 65 classes, 18 object properties, 11 data properties, and $1,167$ individuals. When comparing the creation times (in seconds) of domain experts for rule groups covering 10 different financial events, creating lexico-syntactic rules took $5,839$ seconds until F_1 scores reached 50%, whereas for HIEL rules it only took 356 seconds: a speedup of approximately 16 times. For JAPE rules, the creation times for the same task averaged to 806 seconds, which is about 2 times slower than HIEL. When allowing more creation time for HIEL and JAPE rules (up to $5,839$ seconds), the F_1 scores increase to 78.7% (83.9% precision, 74.1% recall) and 68.7% (85.3% precision, 57.5% recall), respectively, with respect to lexico-syntactic rules, which have an F_1 score of 51.9% (54.9% precision and 49.3% recall). In all cases, HIEL rules outperform JAPE and lexico-syntactic rules in terms of F_1 scores and creation times.

4 Conclusions

We have presented HIEL, a lexico-semantic pattern-based rule language which employs ontological elements (concepts) for information extraction. Through the addition of semantics, we obtain more generic, yet more accurate rules than their lexical counterparts. Additionally, the use of ontologies within the patterns promotes sharing of information as well as easy extendability. Initial experiments on a set of Web news articles show that in terms of creation times, HIEL rules outperform lexico-syntactic rules and the state-of-the-art JAPE rules. Moreover, within a fixed amount of time, HIEL rules yield higher F_1 scores than the JAPE and lexico-syntactic rules.

References

[1] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 168–175. ACL, 2002.

[2] Flavius Frasincar, Jethro Borsje, and Leonard Levering. A Semantic Web-Based Approach for Building Personalized News Services. *International Journal of E-Business Research*, 5(3):35–53, 2009.

[3] Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *14th Conference on Computational Linguistics (COLING 1992)*, volume 2, pages 539–545. ACL, 1992.