Learning Semantic Information Extraction Rules from News

Frederik Hogenboom, Wouter IJntema, and Flavius Frasincar
fhogenboom@ese.eur.nl, wouterijntema@gmail.com, and frasincar@ese.eur.nl

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

Due to the increasing amount of data provided by news sources and the user-specific information needs, recently, many news personalization systems have been proposed both in industry and academia. Often, these systems process news data automatically into information, while relying on underlying knowledge bases, containing concepts and their relations for specific domains. Keeping these knowledge bases up-to-date is a time-consuming and tedious process usually performed by knowledge experts.

Information extraction rules are frequently used in automatic information extraction, yet they are usually manually constructed. As it is difficult to efficiently maintain a balance between precision and recall while using a manual approach, we present a genetic programming-based approach for automatically learning semantic information extraction rules from (financial) news that extract events. The approach makes use of our earlier developed Hermes Information Extraction Language (HIEL) [1], of which the basic constructs are captured by Fig. 1. The latter figure shows an example rule that links CEOs to their companies. Lexical and syntactic elements are indicated by white labels, whereas semantic elements (which make use of an ontology) are indicated by shaded labels.

In HIEL, a rule 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, consisting of three components, i.e., a subject, predicate, and an object, where the predicate describes the relation between the subject and the object. The RHS supports sequences of many different features. First, labels on the RHS associate sequences 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. Moreover, wildcards are also supported, allowing for ≥ 0 tokens or exactly 1 token to be skipped. Last, of paramount importance is the support for semantic elements through the use of ontological classes, instances, and relations.

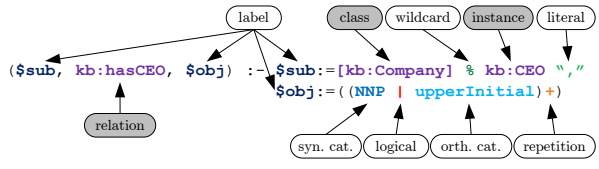

Figure 1: Example HIEL rule

In order to assist domain experts with rule creation, we propose to employ a genetic programming approach to rule learning. HIEL can intuitively be implemented using tree structures and hence fits the required tree structure of the genetic programming operators. Additionally, a genetic programming approach offers transparency in the sense that it gives the user insight into how information extraction rules are learned. Also, a genetic approach often converges to a good solution in a relatively small amount of time.

We implemented our information extraction language and rule learning approach in the Hermes news processing framework (http://people.few.eur.nl/fhogenboom/hermes.html). A preliminary analysis on 500 financial news items shows that, compared to information extraction rules manually constructed by expert users, we are able to find rules that yield a 27% higher F_1-measure after the same amount of rules construction time (5 hours).

In his talk, Frederik Hogenboom, PhD student at the Erasmus University Rotterdam, will be focusing on several aspects of semantic rule learning. First, the specifications of the existing HIEL language are discussed. Second, the genetic programming approach is elaborated on. Third, a preliminary evaluation of the implemented approach is given.

References

[1] Wouter IJntema, Jordy Sangers, Frederik Hogenboom, and Flavius Frasincar. A Lexico-Semantic Pattern Language for Learning Ontology Instances from Text. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 15(1):37–50, 2012.