

## Lexico-semantic Patterns for Information Extraction from Text

The increasing amount of news data has led to many news personalization systems. Often, these systems process data automatically into information, while relying on knowledge bases (ontologies), containing domain-specific concepts and relations. Keeping these ontologies up-to-date is a time consuming and tedious process usually performed by domain experts.

Expert knowledge-driven methods have been a main topic of research for a long time, as large amounts of data are not always readily available, while domain knowledge is usually at hand. These methods require less training data than statistical methods, and their results are more insightful. Various efforts have led to different pattern-languages for information extraction. Most of these are based on lexico-syntactic features, although more lexico-semantic languages are emerging. However, most languages are cumbersome in use, have a limited syntax, and do not make use of domain semantics (expressed in standard languages).

Therefore, we propose a semantics-based pattern language for learning ontology instances from text for knowledge base population. The language makes use of concepts that are defined in an ontology, allowing for inference. Additionally, the developed language supports syntactic categories, orthographical categories, logical operators, and repetition.

We have evaluated our pattern language on a financial data set of 500 news messages on 10 financial events like mergers, profit announcements, etc. for NASDAQ 100 companies. We found that the lexico-semantic patterns are superior to lexico-syntactic patterns in efficiency and effectivity. When applied to news event recognition in the domains of finance and politics, our approach has a precision and recall of approximately 80% and 70%, respectively.