Unsupervised Natural Language Processing

for Knowledge Extraction from Domain-specific Textual Resources

Christian Hänig

Summary

This thesis aims to develop a Relation Extraction algorithm to extract knowledge out of automotive data. While most approaches to Relation Extraction are only evaluated on newspaper data dealing with general relations from the business world their applicability to other data sets is not well studied.

Part I of this thesis deals with theoretical foundations of Information Extraction algorithms. Information Extraction can be divided into two subtasks: Entity Detection and Relation Extraction. The detection of entities is very domain-dependent due to terminology, abbreviations and general language use within the given domain. Thus, this task has to be solved for each domain employing thesauri or another type of lexicon. The task of Relation Extraction can be basically approached by pattern-based and kernel-based algorithms. The latter achieve state-of-the-art results on newspaper data and point out the importance of linguistic features. The performance of state-of-the-art algorithms for POS tagging, syntactic parsing and Relation Extraction is analyzed on automotive data. The findings are: supervised methods trained on newspaper corpora do not achieve accurate results when being applied on automotive data. In order to achieve acceptable results, algorithms have to be trained directly on this kind of data. As the manual annotation of data for each language and data type is too costly and inflexible, unsupervised methods are the ones to rely on.

Part II deals with the development of dedicated algorithms for all three essential tasks. Unsupervised POS tagging is a well-studied task and algorithms achieving accurate tagging exist. None of them disambiguates high frequency words. Most high frequency words bear syntactic information and thus, it is very important to differentiate between their different functions. Domain languages contain ambiguous and high frequent words bearing semantic information (e. g. *pump*). In order to improve POS tagging, an algorithm for disambiguation is developed and used to enhance an existing state-of-the-art tagger. This approach is based on context clustering which is used to detect a word type's different syntactic functions.

An approach to unsupervised syntactic parsing is developed in order to suffice the requirements of Relation Extraction. These requirements include high precision results on nominal and prepositional phrases and detection of endocentric and exocentric constructions. *unsuParse* is based on preferred positions of word types within phrases to detect phrase candidates. Iterating the detection of simple phrases successively induces deeper structures.

Syntactic Relation Extraction is an approach exploiting syntactic statistics and text characteristics to extract relations between previously annotated entities. The approach is based on entity distributions given in a corpus and thus, provides a possibility to extend text mining processes to new data in an unsupervised manner.

To conclude, this thesis presents a complete unsupervised workflow for Relation Extraction – except for the highly domain-dependent Entity Detection task – improving performance of each of the involved subtasks compared to state-of-the-art approaches. Furthermore, this work applies Natural Language Processing methods and Relation Extraction approaches to real world data unveiling challenges that do not occur in high quality newspaper corpora.