Language Engineering for Information Extraction

Martin Schierle

Abstract

Accompanied by the cultural development to an information society and knowledge economy and driven by the rapid growth of the World Wide Web, the world's knowledge is evolving fast, and humans are challenged with keeping up.

Despite all efforts on data structuring, a large part of this human knowledge is still hidden behind the ambiguities and fuzziness of natural language. Especially domain language poses new challenges by having specific syntax, terminology and morphology. Companies willing to exploit the information contained in such corpora are often required to build specialized systems instead of being able to rely on off the shelf software libraries and data resources. The engineering of language processing systems is however cumbersome, and the creation of language resources, annotation of training data and composition of modules is often enough rather an art than a science. The scientific field of Language Engineering aims at providing reliable information, approaches and guidelines of how to design, implement, test and evaluate language processing systems.

Language engineering architectures have been a subject of scientific work for the last two decades and aim at building universal systems of easily reusable components. Although current systems offer comprehensive features and rely on an architectural sound basis, there is still little documentation about how to actually build an information extraction application. Selection of modules, methods and resources for a distinct usecase requires a detailed understanding of state of the art technology, application demands and characteristics of the input text. The main assumption underlying this work is the thesis that a new application can only occasionally be created by reusing standard components from different repositories. This work recapitulates existing literature about language resources, processing resources and language engineering architectures to derive a theory about how to engineer a new system for information extraction from a (domain) corpus.

This thesis was initiated by the Daimler AG to prepare and analyze unstructured information as a basis for corporate quality analysis. It is therefore concerned with language engineering in the area of Information Extraction, which targets the detection and extraction of specific facts from textual data. While other work in the field of information extraction is mainly concerned with the extraction of location or person names, this work deals with automotive components, failure symptoms, corrective measures and their relations in arbitrary arity.

The ideas presented in this work will be applied, evaluated and demonstrated on a real world application dealing with quality analysis on automotive domain language. To achieve this goal, the underlying corpus is examined and scientifically characterized, algorithms are picked with respect to the derived requirements and evaluated where necessary. The system comprises language identification, tokenization, spelling correction, part of speech tagging, syntax parsing and a final relation extraction step. The extracted information is used as an input to data mining methods such as an early warning system and a graph based visualization for interactive root cause analysis. It is finally investigated how the unstructured data facilitates those quality analysis methods in comparison to structured data.