The Essence of Patent Text Mining

In order to create text mining tools for Intellectual property text such as patent, we need to understand the patent life cycle, as well as the linguistic characteristics of the text genre. Compared to the news domain in the patent domain the traditional information retrieval problems are accentuated, from specific search requirements to the linguistic characteristics of the text domain itself. One objective of the thesis is to demonstrate how the impact of the linguistic diversity and information needs have on state-of-the-art text mining techniques. Within the scope of this thesis we develop and modified several text mining tools using state-of-the-art techniques, in order to examine the tools shortcomings and how we adapted them to better suit the patent text genre, in terms of linguistic challenges and information need requirements.

During the process to reduce the gap between training data (news text) and test data (patent text), it became more and more obvious how important the linguistic aspects of a language and a text genre is for domain-specific text mining. The essence of patent text mining for the English language is the noun phrases composed of multi-word units. Noun phrases convey the majority of domain specific terms and are thereby the core of patent text mining applications. Many state-of-the-art text mining methods implicitly postulate that a single word yields the entire scope of a semantic concept. For many main stream text genres this is a valid assumption. However, when a text genre embodies a majority of domain specific concepts as multi-word units in terms of noun phrases instead as noun phrases composed of single words, the assumption will negatively affect the performance of the text mining tools. The state-of-the-art text mining techniques, as well as natural language processing tools decrease in performance when applied on patent text, since they do not recognize the importance of the complexed English noun phrase. By targeting errors connected to noun phrase identification, as well as shifting the focus from bag-of-words to noun phrases, we improve several patent text mining tools in terms of performance and display how important it is to recognize linguistic diversity existing within a language and its sub languages.

In order to demonstrate the importance to recognize different information needs and linguistic diversity within the patent domain, we develop several real-world text mining applications, from information extraction in terms of domain specific terminology, identification and ontology population, to specific question answering systems. During a series of IR experiments, we developed a complete information retrieval system for patent passage retrieval, which incorporate domain knowledge and linguistic diversity within the patent text genre, and meets the specific requirements of patentability and invalidity search. The system components include query formulation from full patent application, partition of a patent document collection into indices based upon lexico-cohesion information and explicit thematic technical fields. The query formulation component includes features such as noun phrase and domain terminology identification, query expansion with domain specific semantic lexical relations. With the IR experiments we improve each evaluation metric significantly compared to the state-of-the-art for the CLEP-IP 2013 test collection: for PRES@100 by 30% (0.563 from 0.433), for recall@100 by 21% (0.654 from 0.540) and on document MAP at 57% (0.300 from 0.191).

In this thesis we postulate and evince that it is equally significant to recognize the differences within a language as between languages, also for a high-density language such as English when it comes to build domain-specific text mining tools.