The Essence of Patent Text Mining

In order to create support tools for patent experts, we need to understand their daily work tasks, as well as the linguistic characteristics of the text genre. Compared to, for instance, the news domain, in the patent domain all types of issues are accentuated, from specific search requirements to the characteristics of the text domain. The main object of the thesis is to give an overview of the patent text domain in order to demonstrate the diversity of linguistic characteristics, as well as information needs that exist in the patent domain. Within the scope of this thesis I develop several text mining applications using state-of-the-art methods in order to examining their shortcomings and how to adapt them to the patent text genre.

It was during the process to reduce the gap between training data (news text) and test data (patent text) that I was confronted with the essence of patent text mining for the English language 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 text genres this is a valid assumption. However, when a text genre embodies a majority of domains specific concepts as multi word units in terms of noun phrases instead as single word noun phrases, the assumption will negatively affect the performance of the applications. The state-of-the-art text mining applications, as well as natural language processing tools decrease in performance when applied on patent text, since they do not recognise the importance of English multi word noun phrases. By targeting errors connected to noun phrase identification, as well as shifting the focus from bag-of-words to noun phrases, I improve several text mining applications presented in the thesis.

In order to demonstrate the importance of multi word noun phrases, I 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. All of the text mining applications come together in the final experiment of the thesis in order to create an advance information retrieval system for patent passage retrieval. The system includes components such as query formulation from full patent documents, noun phrase and domain terminology identification, query expansion with domain specific hyponymy relations based upon distributional semantic information and lexico-syntactic patterns.