Thesis abstract

Thesis title: Low-Bias Extraction of Domain-Specific Concepts
Author: Axel-Cyrille Ngonga Ngomo

Abstract

The recent availability of domain-specific knowledge models in various forms has led to the development of information systems specialized on complex domains such as bio-medecine, tourism and chemistry. Domain-specific information systems rely on domain knowledge in forms such as terminologies, taxonomies and ontologies to represent, analyze, structure and retrieve information. While this integrated knowledge boosts the accuracy of domain-specific information systems, modeling domain-specific knowledge manually remains a challenging task. Therefore, considerable effort is being invested in developing techniques for the extraction of domain-specific knowledge from various resources in a semi-automatic fashion. Domain-specific text corpora are widely used for this purpose.

Most of the current approaches to the extraction of domain-specific knowledge in the form of terminologies or ontologies are limited in their portability to other domains and languages. The limitations result from the knowledge-rich paradigm followed by these approaches, i.e., from them demanding hand-crafted domain-specific and language-specific knowledge as input. Due to these constraints, domain-specific information systems exist currently for a limited number of domains for which domain-specific knowledge models are available. An approach to remedy the high human costs linked with the modeling of domain-specific knowledge is the use of low-bias, i.e., knowledge-poor and unsupervised approaches. They require little human effort but more computational power to achieve the same goals as their hand-crafted counterparts.

In this work, we propose the use of low-bias approaches for the extraction of domain-specific terminology and concepts from text. Especially, we study the low-bias extraction of concepts out of text using a combination of metrics for domain-specific multi-word units and graph clustering techniques. The input for this approach consists exclusively of a domain-specific text corpus. We use a novel metric, the Smoothed Relative Expectation, to extract domain-specific multi-word units from the input data set. In a second step, a novel binary clustering algorithm called SIGNUM is introduced and applied to the results of the metric. By these means, we compute a domain-specific lexicon. Finally, we use second-order collocations to extract the semantic features of the domain-specific terms contained in the domain-specific lexicon. These terms are then clustered to concepts using the third innovation of this work, the graph clustering algorithm BorderFlow. Our approach is unsupervised and makes no use of a-priori knowledge on language-specific patterns and the like. Therefore, it can be applied to virtually all domains and languages.

We evaluate our approach on two domain-specific data sets from the bio-medical domain against domain-specific terminologies and standard clustering techniques. Overall, we show that low-bias approaches can be used to extract domain-specific concepts of high purity.