SIGNUM
A graph algorithm for terminology extraction

Axel-Cyrille Ngonga Ngomo[1]

University of Leipzig, Johannisgasse 26, Leipzig D-04103, Germany,
ngonga@informatik.uni-leipzig.de,
WWW home page: http://bis.uni-leipzig.de/AxelNgonga

Abstract. Terminology extraction is an essential step in several fields of natural language processing such as dictionary and ontology extraction. In this paper, we present a novel graph-based approach to terminology extraction. We use SIGNUM, a general purpose graph-based algorithm for binary clustering on directed weighted graphs generated using a metric for multi-word extraction. Our approach is totally knowledge-free and can thus be used on corpora written in any language. Furthermore it is unsupervised, making it suitable for use by non-experts. Our approach is evaluated on the TREC-9 corpus for filtering against the MESH and the UMLS vocabularies.

1 Introduction

Terminology extraction is an essential step in many fields of natural language processing, especially when processing domain-specific corpora. Current algorithms for terminology extraction are most commonly knowledge-driven, using differential analysis and statistical measures for the extraction of domain specific termini. These methods work well, when a large, well-balance reference corpus for the language to process exists. Yet such datasets exist only for a few of the more than 6,000 languages currently in use on the planet. The need is thus for knowledge-free approaches to terminology extraction. In this work, we propose the use of a graph-based clustering algorithm on graphs generated using techniques for the extraction of multi-word units (MWUs). After presenting work related to MWU extraction, we present the metric for MWU extraction used: SRE. This metric is used to generate a directed graph on which SIGNUM is utilized. We present the results achieved using several graph configurations and sizes and show that SIGNUM improves terminology extraction. In order to evaluate our approach, we used the Medical Subject Headings (MESH), with which the TREC-9 collection was tagged, and the Unified Medical Language System (UMLS) vocabularies as gold standards. Last, we discuss some further possible applications of SIGNUM and the results generated using it.