A Semi-Automatically Enriched Multi-Lingual Terminology in Commercial Products

P. Daumke[1], M. Poprat[1], D. Milward[2], I. Lewin[2]

[1]Averbis GmbH, Tennenbacher Straße 11, D-79106 Freiburg
[2]Linguamatics Ltd, 324 Science Park, Milton Road, Cambridge CB4 0WG, U.K.

One way to exploit the CLEF-ER challenge results is to semi-automatically enrich the multi-lingual terminology provided to the CLEF-ER participants. In the current version, English is the predominant language (1.8 m synonyms in 531k concepts). Synonyms in other languages are clearly underrepresented (Spanish: 643k, French: 127k, German: 119k and Dutch: 116k). Two leading text mining companies, Averbis and Linguamatics, will show how they plan to incorporate the current version of the terminology in their products and how an enriched and well-balanced terminology will improve their commercial applications in the future.

Averbis offers "Patent Analytics"[1], an innovative and powerful patent data and analytics software. Within this platform, large collections of patents can be classified semi-automatically and accessed by faceted browsing, semantic full-text search as well as navigated based on application-specific terminologies. However, the software in its current version runs only in mono-lingual mode, which is a clear disadvantage when dealing with patents written in various languages. Averbis will tackle this problem by implementing the CLEF-ER multi-lingual terminology. In a first step, patents from the multi-lingual MAREC[2] corpus will be annotated with terms from the terminology. Given the mapping from the terms to UMLS semantic types and groups[3] such as "Anatomy", "Chemicals & Drugs", "Disorders" etc., we can use this information to build facets. Furthermore, the identifiers of annotated terms will be used as features for the classification of documents according to IPC[4] across languages. Given a model that is purely trained on term IDs from annotated English documents for instance, it can also be applied to classify all non-English documents.

Linguamatics will demonstrate the added value of an enriched multi-lingual terminology within their text-mining product, I2E[5]. I2E supports highly scalable real-time document search and fact extraction, using natural language processing as a core underpinning technology, together with plug-in domain knowledge and terminology facilities. Real-time queries allow a generate-and-test development strategy. For example, an initial query based simply on word occurrence will often deliver noisy results but, given a fast enough test cycle, the query can be refined in a number of ways beyond "add more words". Queries can be restricted to parts of the known document structure, or by linguistic structures (e.g. to types of phase), or by confidence thresholding. We will present some of the challenges and some of the solutions to interactive search using a plug-in multi-lingual terminology and multi-lingual documents.

[1] http://www.averbis.de/en/products/patent_analytics
[2] http://www.ir-facility.org/prototypes/marec
[3] http://semanticnetwork.nlm.nih.gov/SemGroups/
[4] http://www.wipo.int/classifications/ipc
[5] http://www.linguamatics.com/welcome/software/i2e