Automatic Structuring of Terminologies as Conceptual Units

Galia Angelova
Institute for Parallel Processing, Bulgarian Academy of Sciences[1]
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria, e-mail <galia@lml.bas.bg>

Terms have a crucial role in technical domains as they enable the communication of domain knowledge. More theoretic disciplines like Linguistics (LSP, Terminology, Terminography) and Information Sciences study the nature and structuring of terms, while other disciplines like Natural Language Processing, Artificial Intelligence, and Knowledge Engineering aim recently at more applied tasks related to the automatic acquisition and conceptualisation of terminological units. These practical approaches consider the terms as linguistic labels of domain-specific concepts. They are increasingly important in the age of Internet, when new terms are coined, used and broadly disseminated every day with a dynamics which prevents the organisation of traditional terminological collections. In fact only the automatic processing can cope with the growing amount of texts and new terms, therefore the advanced reference techniques to terms should be built on top of tools for automatic identification and structuring of terminologies. The present talk summarises recent efforts in this respect.

Automatic term extraction from texts has been a hot research area of Natural Language Processing (NLP) in the last decade. In general, it is implemented by extracting stable collocations (usually noun phrases) from a corpus of technical texts and assigning each unit a score, to measure its importance as a domain term. Advanced projects tackle the multilingual case, as large parallel corpora are publicly available (e.g. the web sites of the European Union contain parallel texts in 20+ natural languages). Successful approaches rely on training data (i.e. preliminary developed training corpora) and apply well-defined evaluation methods. Recently, it is agreed that automatic terminology processing should not only be limited to extracting terms and automatic terminology construction, but has to include also extraction of term descriptions, friend terms, and relations. Mining semantic features of terms usually needs predefined “knowledge patterns” that are manually extracted from i.e. glossaries.

Another field that tackles the problems of automatic terminology extraction and structuring arose recently within the popular paradigm of the semantic web. The area is referred to as ontology acquisition, ontology learning, and/or ontology population. The semantic web approaches are less focused on the precise computational treatment of linguistic facts and target primarily the extraction of terms/concepts and domain-important conceptual relations, such as taxonomic relations between terms, attributes of the concepts denoted by terms, instances of concepts and so on. Important target of these applications is the organisation of terms into ontologies and classification of the instances (i.e. named entities) as concept individuals. Despite the simplicity of some earlier and relatively naïve tools, the field is maturing very quickly and attracts increasing attention. This is due to the fact that the terminology-based ontologies are a must in many advanced areas, for instance they will enable the development (tagging and meta-annotation) of intelligent eLearning content. So the synergy between the NLP and the ontology acquisition efforts is to be foreseen in the near future and it will have a visible impact on the terminological field as well.

[1] The research work reported in this paper is partially supported by the project BIS-21++, funded by the European Commission in FP6 via contract no.: INCO-CT-2005-016639.