GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications

Patrice Lopez

European Patent Office, D-10969 Berlin, Germany
plopez@epo.org

Abstract. Based on state of the art machine learning techniques, GROBID (GeneRation Of BIbliographic Data) performs reliable bibliographic data extractions from scholar articles combined with multi-level term extractions. These two types of extraction present synergies and correspond to complementary descriptions of an article. This tool is viewed as a component for enhancing the existing and the future large repositories of technical and scientific publications.

1 Objectives

The purpose of this demonstration is to show to the digital library community a practical example of the accuracy of current state of the art machine learning techniques applied to information extraction in scholarship articles. The demonstration is based on the web application at the following addresse: http://grobid.no-ip.org.

2 Bibliographical Data Extraction

After the selection of a PDF document, GROBID extracts the bibliographical data corresponding to the header information (title, authors, abstract, etc.) and to each reference (title, authors, journal title, issue, number, etc.). The references are associated to their respective citation contexts. The result of the citation extraction can be exported as a whole or per reference following different formats (BibTeX and TEI) and as COInS[1].

The automatic extraction of bibliographical data is a challenging task because of the high variability of the bibliographical formats and presentations. We have applied Conditional Random Fields to this task following the approach of [1] implemented with the Mallet toolkit [2], based on approx. 1000 training examples for header information, and 1200 training examples for cited references. An evaluation with the reference CORA dataset showed a reliable level of accuracy of 98,6% per header field and 74.9% per complete header instance, 95,7% per citation field and 78.9% per citation instance.

[1] See http://ocoins.info

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 473–474, 2009.

By selecting the *consolidation* option, GROBID sends a request to Crossref web service for each extracted citation. If a core of metadata (such as the main title and the first author) is correctly identified, the system will possibly retrieve the full publishers metadata. These metadata are then used for correcting the extracted fields and for enriching the results. Interestingly, the instance accuracy for citations goes up to 83.2% on the CORA dataset with this option.

3 Multi-level Term Extraction

If the option *term extraction* is selected, the header will be enriched with two lists of terms; a list of disciplinary domains for the purpose of a general categorization of the article and a list of the most significant terms extracted from the whole body of text. In addition, each citation is enriched with a third list of terms extracted from the different citation contexts in order to capture the important discriminant aspects for which the reference is used.

The usage of terms of domain-specific terminologies is admittedly viewed as one of the most distinguishing features of scientific and technical documents. Term extraction in GROBID follows the approach of [3]. A term is characterized by two scores; one representing its *phraseness* (or lexical cohesion), i.e. the degree to which a sequence of words can be considered a phrase, and one representing its *informativeness* , i.e. the degree to which the phrase is representative of a document given a collection of documents. A linguistic chain comprising language identification, sentence segmentation, word tokenization, POS tagging and lemmatization is first applied. Noun phrases are extracted as term candidates. The Dice coefficient is computed for evaluating the *phraseness* of a term. The *informativeness* is evaluated based on the estimation of the deviation between the document and a backgroung HMM language model based on a 18 millions corpus of English Wikipedia articles.

4 Application to Digital Libraries

We believe that the text processing modules will be a central component of the future digital libraries. The goal of GROBID is to support various user tasks in relation to large article repositories, in particular the assistance for self archiving of articles, the pre-processing of documents for information retrieval, the generation of reliable OpenURL links or automatic citation suggestions.

References

1. Peng, F., McCallum, A.: Accurate Information Extraction from Research Papers using Conditional Random Fields. In: Proceedings of HLT-NAACL (2004)
2. McCallum, A., Kachites, A.: MALLET: A Machine Learning for Language Toolkit (2002)
3. Tomokiyo, T., Hurst, M.: A language model approach to keyphrase extraction. In: Proceedings of ACL Workshop on Multiword Expressions (2003)