EχATO LP – a tool for domain relevant terms extraction

Lucelene Lopes, Paulo Fernandes, Renata Vieira, Guilherme Fedrizzi, and Daniel Martins
Faculdade de Informática – FACIN – PUCRS
Porto Alegre – RS – Brazil
{lucelene.lopes, paulo.fernandes, renata.vieira, guilherme.fedrizzi}@pucrs.br

Abstract. This paper presents a software tool to extract relevant terms from Portuguese texts. EχATO LP extracts the most frequent noun phrases in an annotated corpus. The annotation is provided by the PALAVRAS parser. The software tool offers different options to improve the quality of extraction that goes from post-treatment of the parser annotation to application of linguistic and statistical criteria. EχATO LP also offers some additional features to compare extracted terms with reference lists, to compute efficiency numerical indexes and to search for terms in the corpus.

Term extraction from corpora is usually the basis of many Natural Language Processing (NLP) task such as automatic glossary construction [7], text categorization [4] and even ontology learning [3]. Term extraction, as many other NLP applications, can benefit from both linguistic and statistical approaches, as the combination of these two approaches often offers better results than each of the approaches separately.

EχATO LP software tool [6] thus uses both linguistic and statistical approaches to extract and select domain significant terms from a an annotated domain corpus. From a linguistic point of view, the extraction is based on the syntactic annotation performed by the parser PALAVRAS [2]. The candidate terms are terms annotated as noun phrases by the parser according to an extra set of discard and transformation rules. From a statistical point of view, those candidate terms are subject to frequency analysis, i.e., in order to select the more frequent ones.

Figure 1(a) graphically presents the software architecture. The basic input is a set of .xml files which are the files with the annotated texts of the corpus. The extraction process consider a set of discard and transformation rules to, respectively, discard some noun phrases that may be unwanted, e.g., noun phrases with numerals, or to adapt some noun phrases to the purpose of extraction, e.g., remove articles. The user can chose by the tool options which rules of these two sets are to be applied. Figure 1(b) upper screenshot presents the interface where the user can choose from all these options.

Once the candidate terms are extracted, their frequencies in the corpus and in each text is computed. Then, by user choice, some of the candidate terms can be selected according to different criteria, e.g., keeping only the 10% more frequent
terms in all corpus. The lists of extracted terms can be compared with reference lists in order to compute usual evaluation metrics like precision, recall and f-measure. Figure 1(b) lower screenshot presents the interface for this comparison.

ExATOlp is an ongoing project, therefore many future extensions are planned. One of the next extensions planned is to aggregate a bootstrap method to increase the quality of selected terms in a similar way as the work of Baroni and Bernardini [1]. Another planned extension is to compute other types of frequency, e.g., the popular tf-idf [5].

References