EXTRACTING TERMINOLOGY BY LANGUAGE INDEPENDENT METHODS

Sanja Seljan, Ivan Dunđer, Hrvoje Stančić/Zagreb

The paper presents automatic extraction process from monolingual text performed by three language independent tools, but relying on different principles. The research is conducted on the domain of pharmaceutical documentation. After the digitization process and use of OCR techniques, the automatic extraction process is performed. Results are compared with reference terminology list created by responsible institution and evaluated by measures of recall, precision and F-measure. Results are discussed in the frame of possible integration into the process of digital archiving.

Key words: automatic terminology extraction, statistical tools, language independent methods, evaluation, indexing, terminology, digital archiving

1. Introduction

Today's business processes heavily relay on the possibilities of utilizing digital and digitized documents. While digitally born and archived documents could be easily, and in some cases automatically, recognised and classified, large sets of divergent digitized documents are not so easily recognised and classified. Firstly, they have to be processed by OCR solutions and then they have to be, ideally automatically, recognised as certain types or classes of documents. This is relatively easy to accomplish if there are enough distinguishing information, e.g. barcode, uniform heading and subheading structure etc. However, if the document set is comprised of many different kinds of documents, such as the one we have analysed, with scarce layout similarities yet with abundant similarities relevant for the classification terminology analysis could be useful. If this proves possible and efficient, the solutions based on this concept could be integrated into the process of digital archiving.

Automatic extraction of corpus-based terminology can help in building terminology lists which represent valuable resource for the research, education and practical implementation. Specific terminology lists represent an intermediate step between the free text and the controlled vocabulary. Such lists can be used in information retrieval, in document indexing, in machine learning, in education, or extended to cross-language information access. Terminology extraction could be performed on monolingual or bilingual/multilingual texts by various terminology extraction methods relying on statistical or language approaches, or on hybrid models. In this paper term candidates are regarded as n-gram word sequences (Deane, 2005). In order to rank candidate terms, various statistical measures have been used, such as frequency-based filtering (Daille et al., 1994), C-Value (Ananiadou, 1994), NC-Value (Frantzi et al., 2000), log-ikelihood and mutual information (Pantel and Lin, 2001), TFIDF (Basili et al., 2001), Termex

(Sclano and Velardi, 2007), etc. In some approaches several language independent methods can be combined, e.g. combination of log-likelihood comparison method to extract monolingual terminology from the source and target sides of a parallel corpus and then use a Phrase-Based Statistical Machine Translation model to create a bilingual terminology with the extracted monolingual term lists (Haque, Penkale, Way, 2014), or performing monolingual terminology extraction for one language, and then aligning these term candidates to the other language's candidates (Ha, Fernandez, Mitkov, Corpas, 2008), or combining several statistics-based language independent methodologies (Teixeira, Lopes, Ribeiro, 2011 and 2013), or using approach that generates candidate terms directly from the aligned words and phrases in a parallel corpus and then use frequency information in order to determine the term specificity (Lefever, Macken, Hoste, 2009).

Evaluation of extracted terminology candidates requires considerable human expertise in evaluation and final compilation, depending on its purpose and type of the final users. The paper presents research on (semi)automatic terminology extraction process performed by three language independent tools, but relying on different principles. Terminology extraction was performed on monolingual texts, mainly on Croatian, but also on English and German languages. The research was conducted on the specific text domain relating to pharmaceutical documentation consisting of reports, approvals and decisions on chemical and pharmaceutical documentation and instructions of use. Results are compared with reference terminology list created by responsible institution and evaluated by measures of recall, precision and F-measure.