Evaluating automatic terminology extraction methods for the construction of an ontology

Iveth Carreño, Annaïch Le Serrec and Mylène Boudreau
{si.carreno.cruz,annaich.le.serrec,mylene.boudreau@umontreal.ca}
Observatoire de linguistique Sens-Texte (OLST)
Département de linguistique et de traduction, Université de Montréal

1.1. Introduction

The present work is part of a research project called SACOT [*]: Knowledge Capture and Modeling through Semi-Automatic Construction of Ontologies from Texts, whose main goal consist in investigating, developing and validating innovative natural language processing approaches (NLP) to capture knowledge conveyed in texts using domain-specific terminology with the aim of representing this knowledge in ontologies.

Since term extraction is viewed as a first step to select potential candidates which represent concepts in ontologies, one of the key tasks of the SACOT project is the selection or development of an automatic terminology extraction (ATE) system. In order to do so, the work was divided into three major phases: a) selection of a specialised domain and corpus building; b) review of the existing automatic term extraction methods; and c) evaluation of some representative automatic term extraction systems.

1.2. Corpus building

The corpora used for the SACOT project are written in English and deal with domains of interest for military operations and intelligence analysis. The team focused on the compilation of documents dealing with the domain of terrorism as a whole as well as documents relevant to the sub-domains of counterterrorism, tactics and weapons. In all cases, documents whose main subjects were the Improvised Explosive Devices (IEDs) and chemical, biological, radiological, nuclear (CBRN) terrorism were prioritized. The corpus contains 136 documents gathered from Internet Web sites, electronic files as well as printed references (books and manuals), and amounts to about two million words.

1.3. Review of the existing automatic terminology extraction methods

To determine which extraction method better suits the objective of the project, we first launched a review of a wide variety of existing commercial and research automatic term extractors developed for English. A list of 29 systems was compiled. Factual (name and type of software, designer, reference web site, etc.) and functional information (main approach applied, main functionalities, targeted units, etc.) for each extractor was compiled in a database.

[*] Project funded by the Department of National Defence in Canada (W7701-061885). Chief investigator: Alain Auger (Defence Research and Development Canada – Valcartier) Co-researchers: Patrick Drouin and Marie-Claude L'Homme (OLST-Université de Montréal).

A report describing each of the current automatic extraction approaches was then prepared based on the information gathered from the 29 extractors as well as on the literature related to the subject. The description is based on the traditional classification which presents the ATE systems according to following three major categories: linguistic, statistical and hybrid.

1.4. Evaluation of some automatic term extraction systems

In the third phase, we selected one extraction system per category; usually the most representative ones within those who seemed more innovative in terms of extraction techniques. Overall, we chose five systems: one that applies only statistical measures, one that applies only linguistic information and three hybrid systems. The first of these systems applies an innovative specificity measure, the second one uses contextual information to refine its output and the third one implements a machine-learning algorithm.

In order to evaluate the performance of the selected systems, we followed a four-step process. Firstly, a 50,000-word evaluation sub-corpus was built. For term extraction purposes, only the documents dealing with the IEDs sub-domain were taken from the terrorism corpora. Secondly, three terminologists manually identified terms in the IEDs sub-corpus, first in group and then individually, to finally produce a single reference list of terms. During this process, and in order to agree on what type of units should be retained, the team established a conceptual structure of the IED field. The terminologists focused on the nominal units that belonged to one of the eight conceptual categories previously defined (e.g., types of IEDs, IED components, IED effects, etc.).

Thirdly, each extractor was run on the evaluation corpus and five lists of candidate terms were generated. A method to compare the manual reference list to the lists produced by the automatic systems was defined. The first part of the comparison, which was done automatically by means of an algorithm, consisted in separating the candidate terms in three different lists: one showing the candidates common to both the manual and the automatic lists, the second containing the units proposed by the automatic extractor and the third one showing the units proposed by the terminologists which did not appear in the automatically generated list. The second part of the comparison consisted in categorizing the differences between the human and the automatic extraction lists. To conclude the evaluation process, we computed the results of the comparison in terms of precision and recall for each automatic term extractor.

In our talk, we will focus on the criteria for compiling the reference list and give the results of the evaluation of the term extraction systems we selected.