NorMan Extractor: Automatic term extraction from technical manuals

Fernando Muniz and Sandra Maria Aluísio

Department of Computer Sciences, University of São Paulo
Av. Trabalhador São-Carlense, 400, 13560-970 - São Carlos/SP, Brazil
{fernando.muniz@gmail.com,sandra@icmc.usp.br}
http://www.icmc.usp.br

Abstract. In this paper, we present the Norman Extractor, a tool for automatic term extraction in technical manuals, i.e., document containing instructions for installation, operation, use, maintenance, parts list, and support for the effective deployment of equipment, for example. The extraction method employs both linguistic and statistical knowledge.

Key words: Automatic term extraction, instruction manuals

1 Introduction

In Brazil, the National Indicator of Functional Literacy (INAF) index has been computed annually since 2001 to measure the levels of literacy of the Brazilian population. The 2007 report presented a worrying scenario, defining the following levels [1]: (1) 7% are Illiterate: individuals who cannot read words and phrases; (2) 25% are Rudimentary: individuals who can find explicit information in short and familiar texts; (3) 40% are Basic: individuals who are functionally literate; (4) 28% are Advanced: fully literate individuals. Apart from the text length, the major difference between the three levels of literacy is the ability to deal with the linguistic complexity of the text, which refers to the research area of Text Simplification (TS). TS aims at to maximize reading comprehension of written texts through their simplification. Simplification usually involves substituting complex by simpler words and breaking down and changing the syntax of complex, long sentences [2]. Instructional texts, also called procedural texts, consist of a sequence of instructions, written with precision in order to achieve a goal. The quality of this documentation is critical, because the misinterpretation of any instruction can generate damage to equipments or even human victims.

The textual simplification task of instruction manuals involves keeping unaltered technical terms, since they carry important information for the correct use of the equipments. In order to simplify instructional texts using SIMPLIFICA (www.nilc.icmc.usp.br/porsimples/simplifica/), an authoring tool for writers to adapt original texts into simplified texts, we created a tool for automatic extraction of terms, called NorMan Extractor that is presented in this paper. This tool uses linguistic knowledge and also offers the possibility of using a statistical filter. Using this extractor, an author of a specific manual can apply lexical simplification in the manual without altering technical terms.

2 NorMan Extractor

Norman Extractor is a tool implemented using Java. It has a graphical interface that allows the user to select which text instructions must be processed for extraction of candidate terms. In addition, the user has the option of selecting a corpus to activate the statistical filter based on the C-Value [4] measure.

The first step of the extraction method of Norman is the processing of the text by the parser PALAVRAS [3] to obtain syntactic knowledge of the text. From this point, the method identifies the relations of generation and enablement in the text. The generation relation is a relation that appears between two actions and it demonstrates that after the completion of an action "A", the action "B" occurs automatically, i.e. "A" generates "B". The generation relation is identified via three syntactic patterns: (1) verb + "Para" + infinitive verb, (2) "Se" + subjunctive verb, (3) "Para" + phrases. The enablement relation occurs when the conduct of an action "A" does not result in automatic implementation of the action "B" [4]. The enablement relation is identified by the following syntactic patterns: (1) Sequences, (2) condition "antes", (3) condition "depois". When the system identifies the generation or enablement relations, a filter is applied with POS patterns to extract the first terms candidates. These filters are the same ones used in the ExPorTer [5] project. To use the statistical filter, the user should submit a corpus about an specific equipment for NorMan Extractor, which then calculates the C-Value [6] measure, which improves the common statistical measure of frequency of occurrence in the extraction of terms, making it more sensitive to multi-word terms.

Finally, the system offers the option to export the list of candidate terms in txt format, so that it can be used by other tools, such as SIMPLIFICA, in order to simplify instruction manuals. In the PROPOR demonstration, we will perform all steps of the extraction of terms. We outline a script of our demonstration at http://nilc.icmc.usp.br/norman/extractor/demo.html

References

1. INAF. Indicador de Alfabetismo Funcional INAF/Brasil - 2007. Online available at http://www.acaoeducativa.org.br/portal/images/stories/pdfs/inaf2007.pdf
2. Advaith Siddharthan. Syntactic Simplification and Text Cohesion. PhD Thesis. University of Cambridge, 2003.
3. Bick, Eckhard. The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework, Aarhus: Aarhus University Press (preprint version) – dr.phil. thesis, 2000.
4. Delin, J.; Hartley, A.; Paris, C., Scott, D. ; Vander Linden, K. Expressing Procedural Relationships in Multilingual Instructions, Proceedings of the Seventh International Workshop on Natural Language Generation, pp. 61-70, Maine, USA. 1994.
5. Teline, M. F. Avaliação de métodos para extração automática de terminologia de textos em português. Dissertação de Mestrado. ICMC-USP, São Carlos, 2004.
6. Katerina T. Frantzi , Sophia Ananiadou. Automatic Term Recognition using Contextual Cues. In Proceedings of 3rd DELOS Workshop, 1997.