Educational FACILITA: helping users to understand textual content on the Web

Marcelo A. Amancio, Willian M. Watanabe, Arnaldo Candido Jr., Matheus de Oliveira, Thiago A. S. Pardo, Renata P. M. Fortes and Sandra Maria Aluísio

Department of Computer Sciences, University of São Paulo
Av. Trabalhador São-Carlense, 400, 13560-970 - São Carlos/SP, Brazil
{amancio,watinha,arnaldoc,taspardo,renata,sandra}@icmc.usp.br,matheusol@grad.icmc.usp.br

Abstract. In this paper, we present a solution combining a Web technology and Natural Language Processing (NLP) systems to adapt web content for poor literacy readers. Particularly, we explore the NLP tasks of lexical elaboration and named entity labeling.

1 Introduction

Although the Internet is an essential source of information, which has enabled the digital inclusion of people, it is still a barrier for those with reading disabilities or those in the process of language learning. In Brazil, according to INAF (National Indicator of Functional Literacy), a large number of people belong to the so called rudimentary and basic literacy levels. These people are only able to find explicit information in short texts (rudimentary level) and also process slightly longer texts and make simple inferences (basic level). The 2007 report presented a worrying scenario: 7% of the individuals were classified as illiterate; 25% as literate at the rudimentary level; 40% as literate at the basic level; and only 28% as literate at the advanced level [3]. Thus, we argue that an assistive technology for adapting web content is an urgent necessity for digital inclusion of low literacy people. In this scenario, we present Educational FACILITA (http://vinho.intermidia .icmc.usp.br/watinha/Educational-Facilita/), a Web technology and NLP systems combined to adapt web content for poor literacy readers. Text adaptation is a very well known practice used in educational settings. [1] mentions two different techniques for text adaptation: Text Simplification and Text Elaboration. The first can be defined as any task that reduces the lexical or syntactic complexity of a text while trying to preserve meaning and information and the second aims at clarifying and explaining information and making connections explicit in a text, using definitions, synonyms or hyperonyms of the text words. Educational FACILITA is part of a bigger project on Text Simplification for Brazilian Portuguese language [2].

2 Lexical Elaboration through synonyms

The first part of the lexical elaboration consists in tokenizing the original text and marking the words that are considered complex. The heuristics for find difficult words were to look up in dictionaries of simple words. If a word is not in the simple words dictionary and is not a proper noun, we assume it to be a complex word. Part of this task includes disambiguate words. The last step of the process consists in providing simpler synonyms for the marked words. For this task, we use the thesauruses TeP 2.0 (http://www.nilc.icmc.usp.br/tep2/) and PAPEL for Portuguese language (http://www.linguateca.pt/PAPEL/). This operation is carried out when the user clicks on a marked word. It triggers a search in the thesauruses for synonymous words that are also in the simple words dictionary. If simpler words are found, they are listed in order, from the simpler to the more complex ones. To determine this order, we used Google API to search each word in the web: we assume that the more a word happens, the simpler it is.

3 Named entities recognition and pos-classification

Named Entities (NEs) primarily refer to proper names and targets names of persons, locations, and organizations, which are very often the answers to the common ”W questions” Who? and Where? For this task we adapted one of the open source NE recognition system: Rembrandt (xldb.fc.ul.pt/wiki/Rembrandt). Its 9 general classes and 47 subcategorized subclasses were changed for our target users. The first adaptation was the redefinition of the classes, in order to remove complex subcategories and to change difficult class names by more common ones. These changes make the NE labels more easily understandable by low-literacy readers. Identifying NEs allowed us to obtain short definitions for them in the Wikipedia.

4 Educational FACILITA prototype

In this demo, we present the current version of the Educational FACILITA. First we will show the installation steps. After that, one running example. We outline a script of our demonstration at http://nilc.icmc.usp.br/˜marcelo/EducationalFaci lita/demo.htm

References

1. D. N. Young. Linguistic simplification of sl reading material: effective instructional practice. The Modern Language Journal, 83(3):350-66, 1999.
2. Alusio S., Specia L., Pardo T., Maziero E., Fortes R. Towards Brazilian Portuguese Automatic Text Simplification Systems. In the Proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008), pp. 240-248.
3. INAF (2007). Indicador de Alfabetismo Funcional INAF/Brasil - 2007. Available online at http://www.acaoeducativa.org.br/portal/images/stories/pdfs/inaf2007.pdf