A BOOSTRAPPING SYSTEM FOR DICTIONARY MANAGEMENT AND PARSING

MIHAI ALEX MORUZ[1,2], DAN CRISTEA[1,2]

[1] Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iasi,
[2] Institute of Computer Science, Romanian Academy, Iasi Branch,
mmoruz@info.uaic.ro
dcristea@info.uaic.ro

Abstract

Electronic lexical resources are the base of various NLP tools (POS taggers, chunkers, NERs, etc.). Out of the largest Romanian language dictionary, "Dicționarul Limbii Române" (DLR), a first XML version was obtained during the eDTLR project. Dictionary Entry Parsing is the process through which structured formats are obtained out of complex dictionary entries presented in input in electronic text formats. At the moment, the electronic version of the DLR, the eDTLR, is undergoing a complete reparsing, based on new input data, which will enhance the quality and remove possible missing parts. This paper describes an approach based on a Content Management System for managing parsed entries, both during the reparsing process and after, for public exposure. The system will manage the current status of all recognized entries of a dictionary (unparsed, partially parsed, successfully parsed and parsed with errors), will be capable of calling successive steps of the parsing process (in order to process new volumes or to reparse entries) and will offer a secure way in which researchers can search for entries, create complex queries and view custom statistics.

Key words — Dictionary Entry Parsing, Dictionary Management System, Content Management System

1. Introduction

During recent years, the number and complexity of Natural Language Processing tools, as well as the need for such tools, have steadily increased. Because of this, the requirement for linguistic resources has increased at a similar rate, bringing to the forefront the issue of digitizing existing resources, in order to make them compatible with the tools and processes currently in use.

One such very large linguistic resource is made up of printed dictionaries, built over the course of many years, by large teams of researchers, and generally edited in an unstructured format. The largest such dictionary for the Romanian language is the "Romanian Language Thesaurus Dictionary" (*Dicționarul Limbii Române* – DLR), made up of more than 175.000 words and variants, written over a span of more than one hundred years, by hundreds of lexicographers. The usefulness of this dictionary for the field of NLP is unquestionable, as it is the most detailed dictionary for the Romanian language to date, describing not only word senses (which are represented in a tree-like logical structure), but also assigning a large number of examples for each sense (sentences containing the word sense in question extracted from a large set of

AUTOMATIC MERGING OF MARKED UP TEXTS FOR DICTIONARY ENTRY PARSING

MIHAI ALEX MORUZ[1,2]

[1] *"Alexandru Ioan Cuza" University of Iasi, Faculty of Computer Science*

[2] *Institute of Computer Science, Romanian Academy, Iasi Branch*

mmoruz@info.uaic.ro

Abstract

Dictionary Entry Parsing is the process by which dictionaries, given in an unstructured rich text format are transformed into structured information. This transformation attempts to extract such information as title word, gloss, examples, bibliographical information, etc. Such structured information is a vital step towards the indexing of dictionary content, which would allow for complex interrogation and usage.

Several large dictionaries have already been transformed into such a format – DEX (Explicative Dictionary for Romanian), TLFI (Tresor de la Langue Francaise), DWB (Deutsches Woerterbuch), etc., but most of this work has been done manually, at great expense.

One of the more successful automatic approaches has been that employed for the transformation of the DLR (Thesaurus Dictionary for Romanian) implemented in the eDTLR project (Cristea et al., 2007). The approach proposed, described in detail in (Curteanu et al., 2008), was successfully employed for the Romanian, French, German and Russsian Thesauri, proving that given a properly formatted and represented input text, the precision of the parser exceeds 90%. However, this is entirely dependent on an electronic format that is correct in both form and content.

The precision for the parsing of DLR was low because the input RTF text was obtained by means of OCR, which generated significant errors in formatting and content. While many of the content errors were manually corrected, most of the formatting errors remained, which greatly hindered the parsing process. More recently, the scanning and OCR process for the DLR has been undertaken again, with much better results in terms of formatting but without any manual corrections.

In this paper we propose an automatic method of merging the two versions of the marked up text representing DLR entries by keeping most of the manually corrected text (where it exists) and also keeping the formatting information obtained after the more detailed scanning and OCR, while also guaranteeing the correctness of the XML format thus obtained.

Key words — Dictionary Entry Parsing, Markup Language, XML merging

1. Introduction

The Romanian Thesaurus Dictionary, created by the Romanian Academy, has been published in two series: the Academy Dictionary (DA), 1913-1949 (A-C, D-De, F-K, L-