Using Apache UIMA annotators for Brazilian Portuguese

William Daniel Colen de Moura Silva*[1], Marcelo Finger[1], and Carlos Eduardo Dantas de Menezes[2]

[1] Departamento de Ciência da Computação, IME/USP, Rua do Matão 1010, CEP 055080-090, São Paulo SP, Brasil

[2] Faculdade de Tecnologia e Ciências Exatas, Universidade São Judas Tadeu, Rua Taquari 546, CEP 03166-000, São Paulo SP, Brasil

Abstract. The demonstration shows how to create applications using CoGrOO's – *Corretor Gramatical para o OpenOffice.org* (Grammar Checker for OpenOffice.org) – annotators, including Sentence Detector, Tokenizer, Proper Name Finder, POS Tagger, Chunker and Shallow Parser. The annotators are compliant with the OASIS Standard platform UIMA – Unstructured Information Management Architecture.

1 Introduction

CoGrOO[1] is an open source grammar checker capable of identifying Brazilian Portuguese language mistakes like: pronoun placement; nominal agreement; subject-verb agreement; usage of grave accent (*crase*) employed to indicate the coalescence of preposition "a" (to) plus definite feminine singular determiner "a", yielding "à" and nominal and verbal government.

Aiming at reusability, CoGrOO's annotators are compliant with the Apache UIMA open architecture[3] and are available to the community in the form of free software.

2 CoGrOO UIMA annotators

The currently available annotators are described below:

1. Sentence Boundary Detector: receives a text and splits it into sentences;
2. Tokenizer: receives a sentence and splits it into words and punctuation marks;
3. Named Finder: receives the sentence tokens and identifies the potential proper nouns, such as person names, places and organization names;
4. Part-of-Speech Tagger: receives a sentence and assigns the most probable morphological tag to its lexical items, according to their context;
5. Chunker: receives a tagged text and finds some small noun phrases (NP) and small verbal phrases (VP);

* Scholarship holder CNPq – Brazil

6. Subject-Verb Finder: receives a tagged sentence with NPs and VPs and searches for the subject. If it is found, it marks the NP as a subject of a VP;
7. Grammar Error Detector: this module looks for grammar errors in the input sentence. It is activated after all the previous sentence analysis steps.

3 The Apache UIMA

The Apache UIMA[3] (Unstructured Information Management Architecture) project offers reference architecture for NLP systems and a framework to help its realization. It defines *SOFA* (Subject of Analysis) as any unstructured document, like text documents in natural language, voice records etc. SOFAs are attached to a *CAS* (Common Analysis Structure), which holds the context of the analysis, including the annotations provided by an annotator for that SOFA, and the Type System, which provides information about the annotations types and the input and output of the UIMA components.

An *annotator* analyzes documents and outputs annotations on the document's context. This annotator plus a set of metadata form an *AE* (Analysis Engine), which contains the framework-provided infrastructure. This allows for the easy combination of an AE with other AEs in different flows and deployments.

4 The Demonstration

The following topics will be demonstrated:

1. How to install Apache UIMA;
2. How to use Apache UIMA tools to understand CoGrOO UIMA capabilities;
3. How to create a simple application using CoGrOO UIMA modules: it will be offered a prototype of an opinion analyzer[2], based on morphological information extracted from input text and a semantic dictionary.

The requirements to reproduce the demonstration are Java 1.6, Apache UIMA framework 2.2.2 or better[4] and CoGrOO UIMA annotators[5].

References

1. J. Kinoshita, L.N. Salvador, C.E.D. Menezes, and W.D.C. Silva. Cogroo - an openoffice grammar checker. In *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on*, oct 2007.
2. P. R. Pasqualotti. Recognition of emotion expressions in computer-mediated interaction (in portuguese). Master's thesis, Unisinos, 2008.
3. W.D.C. Silva, M. Finger, and C.E.D. Menezes. Open text annotators using apache uima. In *PROPOR 2010*, April 2010.

[3] Apache UIMA, `http://incubator.apache.org`
[4] Download Apache UIMA Framework and SDK, `http://incubator.apache.org/uima/downloads.cgi`
[5] CoGrOO UIMA, `http://ccsl.ime.usp.br/cogroo`