Whole-Part Relations Rule-Based Automatic Identification: Issues from Fine-Grained Error Analysis

Ilia Markov¹, Nuno Mamede²,³, and Jorge Baptista³,⁴

¹ Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN)  
México D.F., Mexico  
markovilya@yahoo.com

² Universidade do Algarve/FCHS and CECL  
Faro, Portugal  
jbaptis@ualg.pt

³ INESC-ID Lisboa/L2F – Spoken Language Lab  
Lisboa, Portugal  
{Nuno.Mamede,jbaptis}@L2F.INESC-ID.PT

⁴ Universidade de Lisboa/IST  
Lisboa, Portugal  
Nuno.Mamede@ist.utl.pt

Abstract. In this paper, we focus on the most frequent errors that occurred during the implementation of a rule-based module for semantic relations extraction, which has been integrated in STRING, a hybrid statistical and rule-based Natural Language Processing chain for Portuguese. We focus on whole-part relations (meronymy), that is, a semantic relation between an entity that is perceived as a constituent part of another entity, or a member of a set. In this case, we target the type of meronymy involving human entities and body-part nouns. We describe with some detail the decisions that were made in order to overcome the errors produced by the system and the solutions adopted to improve its performance.

Keywords: whole-part relation, meronymy, body-part noun, Portuguese, error analysis.

1 Introduction

Automatic identification of semantic relations contributes to cohesion and coherence of a text and can be useful in several other Natural Language Processing (NLP) tasks such as opinion mining, question answering, text summarization, machine translation, information extraction, information retrieval, and others [10].

The goal of this work is to improve the extraction of semantic relations between textual elements in STRING, a hybrid statistical and rule-based NLP
chain for Portuguese\cite{17}, by targeting the most frequent errors that occurred during the implementation of the whole-part relations extraction module. Whole-part relations (meronymy) is a semantic relation between an entity that is perceived as a constituent part of another entity, or a member of a set. In this case, we focus on the type of meronymy involving human entities and body-part nouns (henceforward, $Nbp$) when they co-occur in texts.

This paper is structured as follows: Section 2 briefly describes related work on whole-part dependencies extraction, while Section 3 explains how this task was implemented in STRING; Section 4 presents the evaluation procedure; Section 5 describes with some detail how the error analysis was carried out; Section 6 illustrates the results of the performance of the system after the error analysis; and Section 7 draws the conclusions.

## 2 Related Work

Meronymy is a complex relation that “should be treated as a collection of relations, not as a single relation” \cite{15}. In NLP, various information extraction techniques have been developed in order to capture whole-part relations from texts.

Hearst \cite{13} tried to find lexical correlates to the hyponymic relations (type-of relations) by searching in unrestricted, domain-independent text for cases where known hyponyms appear in proximity. The author proposed six lexico-syntactic patterns; he then tested the patterns for validity and used them to extract relations from a corpus. To validate his acquisition method, the author compared the results of the algorithm with information found in WordNet \cite{4}. The author reports that when the set of 152 relations that fit the restrictions of the experiment (both the hyponyms and the hypernyms are unmodified) was looked up in WordNet: “180 out of the 226 unique words involved in the relations actually existed in the hierarchy, and 61 out of the 106 feasible relations (i.e., relations in which both terms were already registered in WordNet) were found.” \cite{13} p. 544. The author claims that he tried applying the same technique to meronymy, but without great success.

Girju et al. \cite{10,11} present a supervised, domain independent approach for the automatic detection of whole-part relations in text. The algorithm identifies lexico-syntactic patterns that encode whole-part relations. The authors report an overall average precision of 80.95\% and recall of 75.91\%. The authors also state that they came across a large number of difficulties due to the highly ambiguous nature of syntactic constructions.

Van Hage et al. \cite{12} developed a method for learning whole-part relations from vocabularies and text sources. The authors reported that they were able to acquire 503 whole-part pairs from the AGROVOC Thesaurus\footnote{http://www.fao.org/agrovoc} to learn 91 reliable whole-part patterns. They changed the patterns’ part arguments with known entities to

\footnote{https://string.12f.inesc-id.pt/ [last access: 22/09/2014].}

\footnote{http://www.fao.org/agrovoc [last access: 12.08.2014].}