The Impact of Enriched Linguistic Annotation on the Performance of Extracting Relation Triples

Sanghee Kim, Paul Lewis, and Kirk Martinez

Intelligence, Agents, MultiMedia Group, Department of Electronics and Computer Science,
University of Southampton, U.K.
{sk,phl,km}@ecs.soton.ac.uk

Abstract. A relation extraction system recognises pre-defined relation types between two identified entities from natural language documents. It is important for a task of automatically locating missing instances in knowledge base where the instance is represented as a triple ('entity – relation – entity'). A relation entry specifies a set of rules associated with the syntactic and semantic conditions under which appropriate relations would be extracted. Manually creating such rules requires knowledge from information experts and moreover, it is a time-consuming and error-prone task when the input sentences have little consistency in terms of structures and vocabularies. In this paper, we present an approach for applying a symbolic learning algorithm to sentences in order to automatically induce the extraction rules which then successfully classify a new sentence. The proposed approach takes into account semantic attributes (e.g., semantically close words and named-entities) in generalising common patterns among the sentences which enable the system to cope better with syntactically different but semantically similar sentences. Not only does this increase the number of relations extracted, but it also improves the accuracy in extracting relations by adding features which might not be discovered only with syntactic analysis. Experimental results show that this approach is effective on the sentences of the Web documents obtaining 17% higher precision and 34% higher recall values.

Keywords: relation extraction, information extraction, inductive logic programming

1 Introduction

When organisations (e.g. 'museum' or 'gallery') hold an immense quantity of information in the formats of electronic documents or databases, missing values for some data can occur. Examples are the names of people who participated in the creation of an art work or historical events which influenced the artists. To extract such missing values, we might need to rely on additional information sources, like the Web. The Web exists as the largest information repository and new data are continuously added. The observation that most of the Web documents are free-texts in various structures and vocabularies emphasizes the importance of techniques that can extract a piece of information of interest.