A Hybrid Approach for Relation Extraction Aimed at the Semantic Web

Lucia Specia and Enrico Motta

Knowledge Media Institute & Centre for Research in Computing
The Open University - Walton Hall, MK7 6AA, Milton Keynes, UK
{L.Specia, E.Motta}@open.ac.uk

Abstract. We present an approach for relation extraction from texts aimed to enrich the semantic annotations produced by a semantic web portal. The approach exploits linguistic and empirical strategies, by means of a pipeline method involving processes such as a parser, part-of-speech tagger, named entity recognition system, pattern-based classification and word sense disambiguation models, and resources such as an ontology, knowledge base and lexical databases. With the use of knowledge intensive strategies to process the input data and corpus-based techniques to deal both with unpredicted cases and ambiguity problems, we expect to accurately discover most of the relevant relations for known and new entities, in an automated way.

1 Introduction

Relation Extraction (RE) consists of the identification of the semantic relations between pairs of terms in unstructured or semi-structured natural language documents. Semantic relations are useful for several applications, including the acquisition of terminological data, construction and extension of lexical resources and ontologies, question answering, information retrieval, semantic web annotation, etc.

In this paper we focus on the application of relation extraction to semantically annotate knowledge coming from raw text, as part of a framework aiming to automatically acquire high quality semantic metadata for the Semantic Web. One of the applications developed within this framework is the *KMi Semantic Web Portal*[1] [6], which analyzes data from texts, databases, and knowledge bases, in order to extract semantic knowledge from all of them in an integrated way, also verifying the quality of this knowledge, according to a domain ontology. The extracted knowledge is formalized into OCML and OWL representations[2].

Currently, the knowledge extracted by the semantic web portal from texts comprises mainly occurrences of entities (instances) that already exist in the knowledge base, and their properties also available in that knowledge base or in databases. It also includes occurrences of new entities, as given by a named entity recognition system, according to the possible types of entities in the domain ontology. Thus, already

[1] http://semanticweb.kmi.open.ac.uk:8080/ksw/index.html

[2] Examples of annotations produced by the KMi Semantic Web Portal for newsletters texts are available in http://plainmoor.open.ac.uk:8080/ksw/pages/news.jsp.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 564 – 576, 2006.

existent entities are semantically annotated with their properties provided by the knowledge base and databases. However, new knowledge about entities (especially relational) is not taken into account. Moreover, little is done with new entities, which are annotated only with their types.

In that context, the relation extraction approach presented here aims to identify the semantic relations between entities in the input texts. These include already existent relations between the entities in the knowledge base, new relations predicted as possible by the domain ontology, or completely new (unpredicted) relations. Additionally, new entities are identified in a more comprehensive way, and their relations are also extracted. As a consequence, extra knowledge about (existing and new) entities can be acquired, yielding a richer representation of the input data, and helping to solve problems that arise when mapping this unstructured data into a semantic representation, such as ambiguities. By identifying new entities in the text and recognizing their types, the approach could also be applied to ontology population. Moreover, since it extracts new relations between entities, it could be used as a first step for ontology learning.

The relation extraction approach makes use of a domain ontology, a knowledge base, and lexical databases, along with knowledge-based and empirical resources and strategies for linguistic processing. These include a lemmatizer, syntactic parser, part-of-speech tagger, named entity recognition system, and pattern matching and word sense disambiguation models. The input data used in the experiments with our approach consists of English texts from the Knowledge Media Institute (KMi)[3] newsletters. We believe that by integrating corpus and knowledge-based techniques and using rich linguistic processing strategies in a completely automated and unsupervised fashion, the approach can achieve more effective results than the previous work, in terms of both accuracy and coverage.

In the remaining of this paper we first describe some cognate work on relation extraction, particularly those exploring empirical methods, for various applications (Section 2). We then present our approach, showing its architecture and describing each of its main components (Section 3). Finally, we discuss next steps (Section 4).

2 Related Work

Several approaches have been proposed for the extraction of relations from unstructured sources. Recently, they have focused on the use of supervised or unsupervised corpus-based techniques in order to automate the task. A very common approach is based on pattern matching, with patterns composed by subject-verb-object (SVO) tuples. Interesting work has been done on the unsupervised automatic definition of patterns from a small number of seed patterns. These are used as a starting point to bootstrap the pattern learning process, by means of semantic similarity measures [20, 16].

Most of the approaches for relation extraction rely on the mapping of syntactic dependencies, such as SVO, onto semantic relations, using either pattern matching or other strategies, such as probabilistic parsing for trees augmented with annotations for entities and relations [11], or clustering of semantically similar syntactic dependencies, according to their selectional restrictions [5].

[3] http://kmi.open.ac.uk/