Domain Information for Fine-grained Person Name Categorization

Zornitsa Kozareva, Sonia Vazquez and Andres Montoyo

Departamento de Lenguajes y Sistemas Informaticos
Universidad de Alicante
{zkozareva,svazquez,montoyo}@dlsi.ua.es

Abstract. Named Entity Recognition became the basis of many Natural Language Processing applications. However, the existing coarse-grained named entity recognizers are insufficient for complex applications such as Question Answering, Internet Search engines or Ontology population. In this paper, we propose a domain distribution approach according to which names which occur in the same domains belong to the same fine-grained category. For our study, we generate a relevant domain resource by mapping and ranking the words from the WordNet glosses to their WordNetDomains. This approach allows us to capture the semantic information of the context around the named entity and thus to discover the corresponding fine-grained name category. The presented approach is evaluated with six different person names and it reaches 73% f-score. The obtained results are encouraging and perform significantly better than a majority baseline.

1 Introduction

The Named Entity Recognition (NER) task was first introduced in the Message Understanding Conference (MUC) as it was discovered that most of the elements needed for the template filling processes in Information Extraction systems are related to names of people, organizations, locations, monetary, date, time and percentage expressions.

There are two main paradigms for NER. In the first one, NEs are recognized on the basis of a set of rules and gazetteer lists [6], [1]. The coverage of these systems is very high, however they depend on the knowledge of their human creator, the number of hand-crafted rules and the kind of entries in the gazetteer lists. In addition, NER rule-based systems are domain and language dependent, therefore they need lots of time in order to be developed and to be adapted.

The other paradigm is machine learning (ML) based NER. Given a set of feature vectors characterizing a named entity, a machine learning algorithm learns these properties and then assigns automatically NE categories to unseen entities. These systems are easily adaptable to different domains, they can function with language-independent characteristics [16], [17], however, their main drawback is related to the number of hand-labeled examples from which the ML system