Improving Named Entity Recognition

Promoter: Marie-Francine Moens
Supervisor: Niraj Shrestha

1. Introduction
Named entity recognition (NER) is a task of extracting and classifying information units like persons, locations, time, dates, organization names, etc. [1]. The task involves labelling (proper) nouns with suitable named entity tags, and it is usually treated as a sequence prediction problem. NER is an important pre-processing task in many applications in the fields of information retrieval (IR) and natural language processing like searching Web data, machine translation, question answering, summarization, etc.

There are a number of state-of-the-art NER systems like Stanford NER, Illinois, LingPipe, OpenNLP NameFinder, etc. Though most of them are supposed to give quite good results over English written data [2], the NE recognition is not perfect. [3] reports that the performance of the Stanford NER system applied on Web data decreases by 12%-14%. There are many named entities emerging everyday especially person, location and organization names. These names are very hard to identify by the NER system.

2. Project Goal
The main goal of this proposal is to investigate how we can improve a current state-of-the-art NER system. We will acquire additional knowledge on the named entities from Web sources such as news texts, Wikipedia, Freebase, DBpedia, etc. and develop suitable methods to integrate this knowledge into the NER machine learning algorithm.

3. Work Outline
   a) Literature study: exploring and studying current state-of-the-art NER and its performance, and trying to find its weaknesses.
   b) Designing and implementing an adapted NER system.
   c) Training and evaluating the NER system
   d) Discussing and evaluating the results.

4. Required Skills
The student is interested in natural language processing and machine learning. Good knowledge of programming languages is required. For more information about the thesis please contact: niraj.shrestha@cs.kuleuven.be

5. References