Multilingual Named Entity Recognition

Sivaji Bandyopadhyay
Computer Science & Engineering Department
Jadavpur University
Kolkata, INDIA.
sbandyopadhyay@cse.jdvu.ac.in

Abstract

The computational research aiming at automatically identifying named entities (NE) in texts forms a vast and heterogeneous pool of strategies, techniques and representations from hand-crafted rules towards machine learning approaches. Hand-crafted rule based systems provide good performance at a relatively high system engineering cost. The availability of a large collection of annotated data is the prerequisite for using supervised learning techniques. Semi-supervised and unsupervised learning techniques promise fast deployment for many NE types without the prerequisite of an annotated corpus. The main technique for semi-supervised learning is called bootstrapping and involves a small degree of supervision, such as a set of seeds, for starting the learning process. The typical approach in unsupervised learning is clustering where systems can try to gather NEs from clustered groups based on the similarity of context. The techniques rely on lexical resources (e.g., Wordnet), on lexical patterns and on statistics computed on a large unannotated corpus.

In multilingual named entity recognition (NER), it must be possible to use the same method for many different languages and the extension to new languages must be easy and fast. Person names can be recognized in text through a lookup procedure, by analyzing the local lexical context, by looking at part of a sequence of candidate words that is a known name component etc. Some organization names can be identified by looking at contain organization-specific candidate words. Identification of place names necessarily involves lookup against a gazetteer, as most context markers are too weak and ambiguous.

An important feature in multilingual person name detection is that the same person can be referred to by different name variants. The main reasons for these variations are: the reuse of name parts to avoid repetition, morphological variants such as the added suffixes, spelling mistakes, adaptation of names to local spelling rules, transliteration differences due to different transliteration rules or different target languages etc.. Name variants can be found within the same language documents.

The major challenges for looking up place names in a multilingual gazetteer are the following: place names are frequently homographic with common words or with person names, presence of a number of exonyms (foreign language equivalences), endonyms (local variants) and historical variants for many place names etc..

Application of NER to multilingual document sets helps to find more and more accurate informa-tion on each NE, while at the same time rich in-formation about NEs is helpful and can even be a crucial ingredient for text analysis applications that cross the language barrier.

Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 3–4,
Hyderabad, India, January 2008.