Machine Learning Methods for Named Entity Recognition with Limited Resource

Sujan Kumar Saha

Abstract

A named entity (NE) denotes a noun or noun phrase referring to a name belonging to a predefined category like person, location and organization. The task of identifying and categorizing named entities from text is known as named entity recognition (NER). NEs are often the pivotal as well as the most information-bearing elements of a text, and NER systems find application in a number of tasks like information extraction, text mining and machine translation. Due to its immense importance, a substantial amount of work has been carried out for NER system development. Although a lot of work was done in the early 90's to develop NER systems, it is still an active research area. The need for NER systems in various languages and domains, and the language and domain specific difficulties, keep the task alive and interesting.

Machine learning techniques have been extensively used for the development of NER systems. These techniques require a considerable amount of resources, the most important of which is named entity annotated data for training. The preparation of NER resources is costly and time consuming. Although substantial named entity annotated corpora have been created in English and a few other languages, NER resources have not been developed or are not publicly available in most languages. The major challenge in the development of a NER system in such resource poor languages arises from the resource scarcity.

Approaches that are commonly used for the NER system development in resource rich scenario, might not work well when the resources are limited. We have faced such challenges when we have attempted to develop a NER system in Hindi, which is the most important language in India. The development of a NER system in a resource poor scenario can be helped by using some special techniques to handle the difficulties arising from the resource scarcity. In this thesis we have proposed a few such techniques and applied these to the NER task.

Manual annotation of a sufficiently large training corpus is costly and time consuming. Also the additional resources like gazetteer lists and context patterns are useful in the NER task but manual creation of such resources is also time consuming. We have studied a few techniques using which the NER resources can be prepared with reduced manual effort. When a large number of features are applied on an insufficient training data, it causes overfitting and performance degradation. We have proposed a few feature reduction techniques for reducing overfitting in NER. We have achieved performance improvement when the reduced features are used by replacing the original features. In the NER task most of the features are string based. The performance of a NER classifier depends on the measure of similarity between these string feature values. We have defined a NER task specific distance function, which is able to capture the semantic similarity existing between the string feature values. This distance function is used as a kernel in a support vector machine based classifier. We have also worked on semi-supervised learning techniques to make use of a large unannotated corpus to supplement a small sized annotated corpus. Through extensive experiments we have shown the usefulness of

the proposed techniques. Although the proposed techniques are defined in the context of the Hindi NER task, the techniques are quite general and can be applied to other NER tasks as well as related tasks. Hence to test the generalizability of the techniques we have also applied some of these methods to the NER task in the biomedical domain (English). Finally we have attempted to prepare a NER system in another Indian language (Bengali) where there is no availability of annotated data or other resources. With the help of the existing NER system in a related language (Hindi) and several resource light techniques we have been able to develop a Bengali NER system with moderate performance in the no-resource scenario.