Resolution of Data Sparseness in Named Entity Recognition using Hierarchical Features and Feature Relaxation Principle

ZHOU GuoDong, SU Jian, YANG LingPeng

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{zhougd, sujian, lpyang}@i2r.a-star.edu.sg

Abstract. This paper introduces a Mutual Information Independence Model (MIIM) and proposes a feature relaxation principle to resolve the data sparseness problem in MIIM-based named entity recognition via hierarchical features. In this way, a named entity recognition system with better performance and better portability can be achieved. Evaluation of our system on MUC-6 and MUC-7 English named entity tasks achieves F-measures of 96.1% and 93.7% respectively. It also shows that 20K words of training data would have given the performance of 90 percent with the hierarchical structure in the features compared with 30K words without the hierarchical structure in the features. This suggests that the hierarchical features provide a potential for much better portability.

1 Introduction

Named entity recognition is to identify and classify the entity names that occur in each sentence of a document. Generally, the overall algorithm works by considering sentences one at a time as they occur in the document. It is a critical component for information extraction and an important step for other natural language processing applications, e.g. information retrieval, machine translation and language understanding.

During last decade, named entity recognition has drawn more and more attention from the MUC named entity tasks [1, 2]. Previous approaches are mainly rule-based [3, 4, 5, 6]. However, rule-based approaches lack the ability of coping with the problems of robustness and portability.

The current trend is to use the machine-learning approach, which is more attractive in that it is trainable and adaptable. Representative machine-learning approaches include HMM [7, 8, 9], Maximum Entropy [10, 11, 12], Decision Tree [13], Winnow [14], MEMM [15] and Conditional Random Fields [16]. Among these approaches, the evaluation performance of HMMs, MEMMs and CRFs is higher than those of others. The main reason may be due to its better ability of capturing the locality of phenomena, which indicates names in text. Moreover, HMMs, MEMMs and CRFs seem more and more used in named entity recognition because of the efficiency of the decoding algorithms, e.g. the Viterbi algorithm [17] in HMM. Therefore, this paper will focus in this direction.

This paper introduces a Mutual Information Independence Model (MIIM) and a feature relaxation principle to resolve the data sparseness problem in MIIM-based named entity recognition. Moreover, various features are structured hierarchically to