A General and Multi-lingual Phrase Chunking Model Based on Masking Method

Yu-Chieh Wu[1], Chia-Hui Chang[1], and Yue-Shi Lee[2]

[1] Department of Computer Science and Information Engineering, National Central University,
No.300, Jhong-Da Rd., Jhongli City, Taoyuan County 32001, Taiwan, R.O.C.
bcbb@db.csie.ncu.edu.tw, chia@csie.ncu.edu.tw
[2] Department of Computer Science and Information Engineering, Ming Chuan University,
No.5, De-Ming Rd, Gweishan District, Taoyuan 333, Taiwan, R.O.C.
leeys@mcu.edu.tw

Abstract. Several phrase chunkers have been proposed over the past few years. Some state-of-the-art chunkers achieved better performance via integrating external resources, e.g., parsers and additional training data, or combining multiple learners. However, in many languages and domains, such external materials are not easily available and the combination of multiple learners will increase the cost of training and testing. In this paper, we propose a mask method to improve the chunking accuracy. The experimental results show that our chunker achieves better performance in comparison with other deep parsers and chunkers. For CoNLL-2000 data set, our system achieves 94.12 in F rate. For the base-chunking task, our system reaches 92.95 in F rate. When porting to Chinese, the performance of the base-chunking task is 92.36 in F rate. Also, our chunker is quite efficient. The complete chunking time of a 50K words document is about 50 seconds.

1 Introduction

Automatic text chunking aims to determine non-overlap phrases structures (chunks) in a given sentence. These phrases are non-recursive, i.e., they cannot be included in other chunks [1]. Generally speaking, there are two phrase chunking tasks, including text chunking (shallow parsing) [15], and noun phrase (NP) chuncking [16]. The former aims to find the chunks that perform partial analysis of the syntactic structures in texts [15], while the later aims to identify the initial portions of non-recursive noun phrase, i.e., the first level noun phrase structures of the parsing trees [17] [19]. In this paper, we extend the NP chunking task to arbitrary phrase chunking, i.e., base-chunking. In comparison, shallow parsing extracts not only the first level but also the other level phrase structures of the parsing tree into the flat non-overlap chunks.

Chunk information of a sentence is usually used to present syntactic relations in texts. In many Natural Language Processing (NLP) areas, e.g., chunking-based full parsing [1] [17] [24], clause identification [3] [19], semantic role labeling (SRL) [4], text categorization [15] and machine translation, the phrase structures provide down-stream syntactic features for further analysis. In many cases, an efficient and