Clause Boundary Identification Using Conditional Random Fields

Vijay Sundar Ram. R and Sobha Lalitha Devi

AU-KBC Research Centre,
MIT Campus Anna University,
Chromepet, Chennai -44,
India
(Sobha,sundar)@au-kbc.org

Abstract. This paper discusses about the detection of clause boundaries using a hybrid approach. The Conditional Random fields (CRFs), which have linguistic rules as features, identifies the boundaries initially. The boundary marked is checked for false boundary marking using Error Pattern Analyser. The false boundary markings are re-analysed using linguistic rules. The experiments done with our approach shows encouraging results and are comparable with the other approaches

1 Introduction

The clause identification is one of the shallow parsing tasks, which is important in various NLP applications such as translation, parallel corpora alignment, information extraction, machine translation and text-to-speech. The clause identification task aims at identifying the start and end boundaries of the clauses in a sentence, where clauses are word sequences which contain a subject and a predicate. The subject can be explicit or implied. For the clause identification task we have come up with a hybrid approach, where conditional random fields (CRFs), a machine learning technique and rule-based technique are used. The CRFs module with linguistic rules as features identifies the clause boundaries initially. The erroneous clause boundary detections are identified using an error analyzer and those sentences are processed using the rule-based module.

The clause identification was a shared task in CoNLL 2001. The task of identifying the clause boundaries is non-trivial. More research has been done in this task. The initial approaches to this task were using rule-based technique, which was followed by machine learning and hybrid techniques.

Early experiments in the clause boundary detection are Eva Ejeuhed's basic clause identification system for improving AT&T text to speech system [7], Papergeorgiou's rule based clause boundary system as preprocessing tool for bilingual alignment parallel text [15]. Leffa's rule based system reduces clauses to noun, adjective or an adverb, which was used in English/Portuguese machine translation system [10].