Towards Language-Independent Sentence Boundary Detection

Do-Gil Lee and Hae-Chang Rim

Dept. of Computer Science & Engineering, Korea University
1, 5-ka, Anam-dong, Seongbuk-ku, SEOUL 136-701, Korea
{dglee, rim}@nlp.korea.ac.kr

Abstract. We propose a machine learning approach for language-independent sentence boundary detection. The proposed method requires no heuristic rules and language-specific features, such as Part-of-Speech (POS) information, a list of abbreviations or proper names. With only the language-independent features, we perform experiments on not only an inflectional language but also an agglutinative language, having fairly different characteristics (in this paper, English and Korean, respectively). In addition, we obtain good performances in both languages.

1 Introduction

Sentence boundary detection (SBD) is the first step of natural language processing applications. Most of them, including POS taggers and parsers regard sentence as input. Many researchers have considered the SBD task as an easy one so that they have not stated the accurate algorithms for the job.

A sentence usually ends with a punctuation mark such as '.', '?', or '!'. So, identifying sentence boundaries can be done by comparatively simple heuristic rules. The punctuation marks, however, are not always used as a sentence final. Moreover, some sentences may not have any punctuation mark. For example, output texts from automatic speech recognition (ASR) systems are unpunctuated. Even a phrase or a single word would be a sentence by itself. Therefore, in order to acquire accurate results, it is required more and more complicated rules. Writing such rules are both labour-intensive and time-consuming. For these reasons, recently, various machine learning techniques, such as decision tree, neural network, and maximum entropy, have been successfully applied to the SBD task[1][2].

The previous works have been made mainly on English and European languages such as German and French[1]. Few researches have been done for languages other than Roman-alphabet languages. To our knowledge, a recent SBD work about Korean language is only [3]'s. They take a hybrid method (regular expressions, heuristic rules, and decision tree learning), but cannot be applied to languages other than Korean because language-specific features are used (such as a functional syllable list, a sentence-end syllable list, and POS information).

There are at least three main differences between English and Korean when considering SBD: firstly, Korean has no capitalization information, which is very