Korean Compound Noun Decomposition Using Syllabic Information Only

Seong-Bae Park, Jeong-Ho Chang, and Byoung-Tak Zhang

School of Computer Science and Engineering
Seoul National University
151-744 Seoul, Korea
{sbpark,jhchang,btzhang}@bi.snu.ac.kr

Abstract. The compound nouns are freely composed in Korean, since it is possible to concatenate independent nouns without a postposition. Therefore, the systems that handle compound nouns such as machine translation and information retrieval have to decompose them into single nouns for the further correct analysis of texts. This paper proposes the GECORAM (GEneralized COmbination of Rule-based learning And Memory-based learning) algorithm for Korean compound noun decomposition using only syllabic information. The merit of rule-based learning algorithms is high comprehensibility, but they shows low performance in many application tasks. To tackle this problem, GECORAM combines the rule-based learning and memory-based learning. According to the experimental results, GECORAM shows higher accuracy than rule-based learning or memory-based learning alone.

1 Introduction

The nouns that appear successively without a postposition can be concatenated to compose a compound noun in Korean. Such compound nouns have more contextual information compared to single nouns [15], and play important role in natural language processing. The critical issue in handling compound nouns is that the number of possible compound nouns is infinite. Because all compound nouns can not be listed in the dictionary, it is required to decompose given a compound noun into single nouns.

When a compound noun is composed of n syllables, there are theoretically 2^{n-1} kinds of decompositions. Thus, the easiest way to decompose a compound noun is to take the most plausible one among 2^{n-1} decompositions. Many previous studies have been proposed based on this idea. Shim used composite mutual information trained from about a corpus of 1.1 million word size [15]. Lee et al. considered this task as part-of-speech tagging, and applied a Markov model [11].

The main drawback of such statistics-based methods is that it is difficult for human to understand the trained results. On the other hand, the rules, whether they are made manually or automatically, have high comprehensibility. Thus, there have been a number of studies that apply rules to compound noun decomposition. For instance, Kang designed four decomposition rules and two