A Term Weighting Method based on Lexical Chain for Automatic Summarization

Young-In Song, Kyoung-Soo Han, Hae-Chang Rim

Natural Language Processing Lab., Dept. of CSE,
Korea University, Anam-dong 5-ga, Seongbuk-gu, 136-701, Seoul, Korea
{sprabbit, kshan, rim}@nlp.korea.ac.kr

Abstract. We suggest a new term weighting method based on lexical cohesion in a text. To compute cohesion, we use lexical chain with a new lexical chain disambiguation method considering association between words and characteristics of WordNet. In our experiment, the methods show a better result than traditional term weighting methods such as tf and tf.idf.

1 Introduction

Summarization can be defined as the work of extracting essential part of a document as a readable form and deleting some redundant information. From this point of view, finding thematic terms of a document is one of the important tasks in the automatic summarization, and various term weighting methods such as tf.idf have been used for it.

In this paper, we try to develop a new term weighting scheme considering the lexical cohesion, that is to say, semantic relations (e.g. synonym, antonym, etc.) between words. Our motivation is quite simple but reasonable; Thematic words have more semantic relation than others in a text.

One of the efficient methods to compute the lexical cohesion is to use the lexical chain. The lexical chain can be defined as groups or sequences of semantically related words, and the work of [1] and [2] suggest successful ways to use the lexical chain for the automatic summarization. To use the lexical chain for summarization task, there are two problems to be solved;

1. Ambiguity of the lexical chain: A semantic relation between words is dependent on the meaning of each word. If the ambiguity of word sense cannot be resolved, lexical chain also has an ambiguity.
2. Relevance of the lexical chain: All of lexical chains are not relevant to the topic of the document. A method to distinguish useful chains from useless ones is needed.

In previous works, heuristic methods were used for above two problems, based on the number and kind of semantic relations in a chain[2, 3]. However, they are so simple that they cannot help containing some errors. They do not consider anything about words which make a semantic relation. Thus we suggest a new disambiguation method considering word association and term weighting methods based on it.