Clustering of English-Korean Translation Word Pairs Using Bi-grams

Hanmin Jung[1], Hee-Kwan Koo[2], Won-Kyung Sung[1] and Dong-In Park[1]

[1] NTIS Division, KISTI, Korea
jhm@kisti.re.kr
[2] Practical Information Science, UST, Korea

Abstract. This paper describes a clustering algorithm for Korean translation words automatically extracted from Korean newspapers. Since above 80% of English words appear with abbreviated forms in Korean newspapers, it is necessary to make the clusters of their Korean translation words to easily construct bi-lingual knowledge bases such as dictionaries and translation patterns. As a seed to acquire each translation cluster, we repeat to choose an adequate translation word from a remaining translation set using an extended bi-gram-based binary vector matching until the set becomes empty. We also deal with several phenomena such as transliterations and acronyms during the clustering. Experimental results showed that our algorithm is superior to Dice coefficient and Jaccard coefficient in both determining adequate translation words and clustering translations.

1 Introduction

As information technology develops in recent years, many terminologies are rapidly created and discarded. Newspapers are excellent resources to acquire new-coined terms and to inspect their life cycle [4]. About 90% of terms in Korean newspapers, in particular, are originated from foreign languages such as English and Chinese[1] [1]. Some of them are accompanied by original words in English for readers to easily grasp the meaning, for example, "세계무역기구 (WTO)." However, many English words (about 82% in our test set) appear with abbreviated forms, and translations differ like "아시아태평양경제협력기구," "아시아태평양경제협력체," "아태경제협력체," and "아태경제협력회의" for "APEC; Asia-Pacific Economic Cooperation." Such English abbreviated forms tend to cause word sense ambiguities, for example, "Internet Service Provider," "Information Strategic Planning," and "Image Signal Processor" for "ISP." Newspapers also usually use parentheses to represent a pair of translation pairs, but they are not limited to the pairs. Many extraction errors are caused by the free uses of parentheses such as "모델명 S3C2410 (CPU)"[2] and

[1] E.g., "아펙" is a Korean transliterated word for English "APEC," and "경제" is for Chinese "經濟."

[2] "모델명 S3C2410" = "모델명 (Model No.)" + "S3C2410."

© A. Gelbukh (Editor)
Advances in Natural Language Proccesing
Research in Computing Science 18, 2006, pp. 15-24