International Journal of Innovative
Computing, Information and Control
Volume **5**, Number **2**, February **2009**

 ISSN 1349-4198
pp. 543–554

AUGMENTED MUTUAL INFORMATION FOR MULTI-WORD EXTRACTION

Wen Zhang[1], Taketoshi Yoshida[1], Tu Bao Ho[1] and Xijin Tang[2]

[1]School of Knowledge Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
{ zhangwen; yoshida; bao }@jaist.ac.jp

Institute of Systems Science
[2]Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190, P. R. China
xjtang@amss.ac.cn

Received January 2008; revised May 2008

Abstract. *In order to extract multi-words from documents, mutual information (MI), as a statistical method, is the most popular solution under consideration. However, there are two kinds of deficiencies inherent in MI. One is the problem of unilateral co-occurrence, and the other is rare occurrence problem. To attack these two problems, augmented mutual information (AMI) is proposed in this paper to measure word dependency for multi-word extraction. We prove theoretically that AMI has the capacity to approximate MI to capture the independency of individual words, but it will amplify the significance of dependent individual words which may be possible multi-words. And our experimental results on Chinese multi-word extraction demonstrate that AMI method has superior performance to traditional MI method.*
Keywords: Multi-word extraction, Mutual information, Augmented mutual information, Word dependency

1. **Introduction.** A word is characterized by the company it keeps [1]. That means not only the individual word but also its context should be emphasized for further processing. This simple and direct idea motivates the research on multi-words, which is expected to capture the context information of the individual words. Although multi-word has no satisfactory formal definition, it can be defined as a sequence of two or more consecutive individual words, which is a semantic unit, including steady collocations (e.g. proper nouns, terminologies, etc.) and compound words [2-4,7,10,11,16]. Usually, it is made up of a group of individual words, and its meaning is either changed to be entirely different from (e.g. collocation) or derived from the straight-forward composition of the meanings of its parts (e.g. compound phrase).

Generally speaking, there are mainly two types of methods developed for multi-word extraction. One is the linguistic method, which utilizes the structural properties of phrases and sentences to extract the multi-words from documents [2,3]. The other is the statistical method, based on corpus learning with mutual information for word occurrence pattern discovery [4,5]. There are also some other multi-word extraction methods which combine both linguistic knowledge and statistical computation [6-10]. However, as for the statistical methods for multi-word extraction, most of them employ MI directly, or an adaptation of MI without theoretical proof.