Linguistically Enhanced Collocate Words Model

Siaw Nyuk Hiong, Bali Ranaivo-Malançon, Narayanan Kulathuramaiyer, and Jane Labadin

Faculty of Computer Science and Information Technology
University Malaysia Sarawak
Kuching, Malaysia
ftsm2006@yahoo.com, {mbranaivo,nara,ljane}@fit.unimas.my

Abstract. Bag-of-word (BOW) or fixed size window approach for word extraction in natural language text has ignored text structure and context information. Similarly, word co-occurrence based on linear word proximity has also ignored the linguistic criteria of words. This paper aims to propose a semantic window of word to address the needs to provide a context for capturing the structure and context of word in a sentence for analysis. The semantic window of word has linguistic elements which can be injected for collocate word identification. Selected data has been used as case studies. Quantitative analysis has been conducted as well. The proposed approach is evaluated and compared to sliding window which is the baseline. Semantic window is found to perform better than sliding window for linguistically enhanced collocate word extraction.

Keywords: Semantic dependency parsing, linguistic, collocation, semantic window.

1 Introduction

It has been reviewed that bag-of-word (BOW) approach ignores the context of occurrence of the word and association between words in documents [1]. A lot of text structure and context information are lost with BOW [2,3]. Similarly, the employment of a fixed size window in word co-occurrence identification [1], [4-9] is also lacking in capturing the semantic information of a text. Co-occurrence is a statistical view for related items identification [10]. This approach identified word co-occurrence based on linear word proximimity which actually has ignored the linguistic properties of the words [11]. The linguistically motivated view which defines items as syntactically related is called collocation [10]. Words extracted based on collocation can represent the content more accurately compared to isolated word extraction [12]. Padó and Lapata [13] shared similar view that the gist of a document can be better modeled by syntactically related co-occurrence than the surface co-occurrence of words. Many natural language processing (NLP) has applied collocation for research in text classification [14], topic segmentation [15], text summarization [16], machine translation [17,18], information retrieval [19] and word sense disambiguation [20,21,22]. According to Seretan [23], syntax-based approach has the advantage over

A. Jaafar et al. (Eds.): AIRS 2014, LNCS 8870, pp. 230–243, 2014.