Collocation Extraction in Turkish Texts Using Statistical Methods

Senem Kumova Metin[1] and Bahar Karaoğlan[2]

[1] Izmir University of Economics,
Engineering and Computer Science Faculty, Izmir, Turkey
senem.kumova@ieu.edu.tr
[2] Ege University, International Computing Institute, Izmir, Turkey
bahar.karaoglan@ege.edu.tr

Abstract. Collocation is the combination of words in which words appear together more often than by chance. Since collocations are blocks of meaning, they play an important role in natural language processing applications (word sense disambiguation, part of speech tagging, machine translation, etc). In this study, a corpus of Turkish is subjected to the following statistical techniques: frequency of occurrence, mutual information and hypothesis tests. We have utilized both stemmed and surface form of corpus to explore the effect of stemming in collocation extraction. The techniques are evaluated by recall and precision measures. Chi-square hypothesis test and mutual information methods have produced better results compared to other methods on Turkish corpus. In addition, we have found that a stemmed corpus facilitates discrimination between successful and unsuccessful collocation extraction methods.

Keywords: Collocation, collocation extraction.

1 Introduction

Collocations are conventional word combinations that co-occur together so recurrently that they may not be regarded as random combination of words. The term collocation has been first introduced by an English linguist, J. R. Firth, in the book"Modes of Meaning" [1] in which he states that a word can be understood by the company it keeps and gives some examples to illustrate the notion of collocations. In his further study, he states "Collocations of a given word are statements of the habitual or customary places of that word". Later, Sinclair, a student of Firth, defined collocation as the occurrence of two or more words within a short space of each other in a text [2]. In contrast, Hoey [3] gives a more statistical definition, stating that a collocation is the appearance of two or more lexical items together with a probability that cannot be interpreted as random. In Oxford Collocation Dictionary collocation is defined as the co-occurrence of words to produce natural-sounding speech and writing.

Since collocations are arbitrary and indefinite, they have an important effect on meaning in text and speech. As a result, extracting of collocations supports a

H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (Eds.): IceTAL 2010, LNAI 6233, pp. 238–249, 2010.

wide range of natural language processing applications such as natural language generation, machine translation, word sense disambiguation, part of speech tagging, information retrieval, computational lexicography, corpus linguistic search and in some social studies through language ([4], [5]). In order to serve for this wide range of applications, many different methods of collocation exploration can be found in the literature which can be categorized as: statistical and rule based methods. Rule based methods depend especially on part of speech tagging information. On the other hand, statistical methods (frequency measure, mutual information [6], hypothesis testing, etc.) are based on some kind of frequency measure to extract collocations in a given corpus. Smadja's Xtract [7] and the techniques of Kita et al. [8] and Shimohata et al. [9] are also examples of methods known by the names of researchers.

In this study, we have applied some statistical techniques to extract collocations in Turkish and compared the results using recall and precision measures. We have utilized both stemmed and surface-formed corpora to examine the effect of extensive agglutination in Turkish. Although there are many studies on different languages, including English, Spanish, Russian, Chinese, French, to the best of our knowledge, there is no corresponding study on Turkish in this concept. We believe that our results may further open research in the field of agglutinative languages, especially Turkish.

In section 2, the term "collocation" is presented. In section 3, we have given previous work on Turkish collocations. In section 4, collocation extraction techniques which are implemented in the study are briefly described. In section 5, experimental setup which clarifies utilized corpora, the base set and evaluation method is given. Section 6 involves the implementation results. Finally section 7 deals with the discussion of the above study.

2 Collocation

As it is evident from different definitions of collocation in recent works, there are no known rules for the formation of collocations. Although researchers do not have a total consensus on either the definition of collocation or the rules by which they are created, common features collected from different studies may be listed and defined as in below:

Collocations are recurrent. Of all properties which discriminates collocations from other word combinations, recurrence is the easiest property to measure. As a result, almost all extraction techniques depend on some kind of frequency measure ([4],[6], [7], [10], [11]).

Collocations are arbitrary and language specific. There are no known rules that define which words collocate and how a word chooses a particular word or words from millions of different words in language to create a collocation. For example "strong" is a common collocation with "coffee" in English, But there is no clear explanation for the preference of this word instead of "powerful". Also collocations may change in different languages depending on the social or