Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 433 – 441

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Outlier Detection in Automatic Collocation Extraction

Octavio Santana Suárez[a], Isabel Sánchez-Berriel[b]*, José Pérez Aguiar[a], Virginia Gutiérrez Rodríguez[b]

[a]Data Group and Computational Linguistic (GEDLC), University of Las Palmas de Gran Canaria, Edificio Departamental de Informática y Matemáticas, Campus Universitario de Tafira, Las Palmas de Gran Canaria, 35017, Spain
[b]Department of Computer and Systems, La Laguna University, Facultad de Matemáticas Campus de Anchieta,La Laguna, 38200

Abstract

In this paper we have analysed different association measures between words, generally used for the automatic extraction of collocations in textual corpus. Specifically, they have been considered: relative frequency, mutual information, z-score, t-score and Dunning's test. The volume of handled corpus (300000000 words) requires reviewing of the usual approach to this matter, so a solution that is based on methods used to detect statistical outliers is proposed. It is evident from the results that a lot of free combinations extracted with collocations coming from the comparison of words with very different frequencies of use. For this reason, they are applied considering that each word generates a different sample, instead of generating rankings which come from corpus considered as a single sample. The experiment is also performed on a corpus with a much smaller amount of words and the results are reported so contrasted with those obtained with the full corpus. The conclusions and contributions arising give response automatic extraction of collocations from a textual corpus regardless its volume.

Keywords: collocations, association measures, outliers.

* Corresponding author. Tel.: +34-922319449; fax: +34-.
E-mail address: isanchez@ull.edu.es

doi:10.1016/j.sbspro.2015.07.463

1. Introduction

The term collocation in this work relates to combinations of words used recursively in a language. This definition, from the linguistic point of view is simple, but the phenomenon focuses on recurrence, which allows automatic extraction of them by processing textual corpus. Examples of Spanish collocations are: *tener apetito* (to have an appetite), *afrontar riesgos* (to take risks), *competir duro* (to compete hard), *conversación animada* (animated conversation)… In this problem should be considered the formal flexibility of the elements in the collocation, since they allowed changing grammatical category, adjectival modification, transformation passive, nominalization,… For this reason, we approach the problem from combinations of canonical forms, rather than graphic words: "*el trasplante de órganos*" (organ transplant,), and "*trasplantó el órgano*" (He transplanted the organ) are considered instances of the same collocation(Koike, 2001). However, the characteristic that distinguishes them from other combinations is the preference, as the speakers of the language could choose another combination to convey the meaning intended, but have mostly chosen to use collocations.

The main problem addressed is the automatic extraction of collocations by processing corpus evaluating association measures or collocational indicators that capture the relationship established between the base and the collocative through the use made of both elements in the corpus, individually and together.

Specifically, we consider the corpus as a sample of the use of the language use, any combination is expected to appear in it by chance, i.e. the general case that is considered in the production of combinations is free combinations.

The statistical concept of independence is an ideal tool to determine when two phenomena have occurred together by chance, that is independently. If instead of this, there has been motivation, i.e. the fact that some of them happen in some way influences the possibilities to originate the other. In terms of probabilities, this stated in the statistical law:

$P(x,y) = P(x) * P(y)$

In this paper, "x" means the word x appears in the corpus, same for "y" and "(x, y)" represents the co-occurrence of the word x and word y.

Actually, a textual corpus is a sample of the use of language, so instead of working with probabilities, calculations are done on estimates it through the observed frequencies. In this case we will refer the number of occurrences of the combination and the words x and y individually.

Nomenclature

f(x,y)	*x, y* co-ocurrence frequency
f(x)	word x frequency
f(y)	word y frequency
$\bar{f}$	arithmetic mean of frequencies
s	standard deviation
D-test	Dunning's test

1.1. Association Measures

Based on statistical independence have arisen various proposals to measure the association between two words that appear together in the corpus and use it to make ranking scores that allow to order combinations as collocations candidates. The most simple association measure is the relative frequency, also so called frequency of appearance of *x* with *y* (Koike, 2001). If this value is high, it means that if *x* appears it's probably that also appears *y*, it doesn't meant that we obtain the same value that appearance *y* with *x* frequency.