ISSN 2007-9737

An Analysis of Variance Method for Detection of Collocations in a Pedagogical Domain Corpus

Yuridiana Alemán, María Somodevilla, Darnes Vilariño

Benemérita Universidad Autónoma de Puebla,
Computer Science Faculty,
Mexico

{yuridiana.aleman, mariajsomodevilla, dvilarinoayala}@gmail.com

Abstract. In this paper, an exploratory experiment, based on analysis of variance, was carried out in order to get collocations in a pedagogical domain corpus. A semi-automatic corpus containing learning styles papers in Spanish was built. Afterwards, the corpus was lemmatized and a bigrams representation was extracted. The proposed method consists on divide the list of bigrams in quartiles, and analyzing the variance on each one of them. A list of collocations, which was evaluated using a gold standard built by an expert in the domain, was retrieved from each experiment according to established thresholds for the method. Results showed a retrieved list with important collocation in the selected domain.

Keywords. Pedagogical domain, variance, collocations, ontology, important concepts.

1 Introduction

Natural Language Processing (NLP) defines a collocation as an expression consisting of two or more words that correspond to some conventional way of say things. Collocations have important in NLP such as information retrieval, parsing, and principal concepts detection, among others. Also, collocation extraction is the first step in many specific tasks. Some of the techniques for collocations detection includes semantic metrics [2], ontologies [3], statistical models [5] and the use of external resources depending of the language [4].

In other areas as image preprocessing, there are techniques that can be used in NLP, for example, the method by Otsu assumes that the original image contains pixels from two classes, whose intensity distributions are unknown. The goal is to find a threshold q such that the resulting background and foreground distributions are maximally separated, which means that they are (*a*) each as narrow as possible (have minimal variances) and (*b*) their centers (means) are most distant from each other [1].

This work is focused on the variance analysis to create groups and obtain collocations. The motivation behind the use of a threshold for creating different classes is grounded on the analysis of words pairs frequency. There are words that always appear together but the frequency is low and there are not collocations, the analysis by groups is necessary for deleting the groups with low o null variance.

A pedagogical corpus in Spanish was used for the experiments, afterwards, a simple frequency and conditional probability were calculated; this process lets to divide the words pair in quartiles, and get the variance by group. As evaluation, a principal concept list was matched with the results, as well as a list of pairs validated by an expert in the domain.

This article is organized in five sections described as follows. Section 2 presents related works about collocations and threshold detection for classification. In section 3, experiments are carried out followed by a detailed discussion of results in Section 4. Finally, Section 5 outlines conclusions and future work of the research.

2 Related Work

In [2], a rank aggregation method for collocations detection task was proposed. This method consists of applying some well-known methods such as Dice method, chi-square test, z-test and likelihood ratio. Then, aggregating the resulting collocations rankings by rank distance and Borda score. The proposed aggregation method performs better then each individual method taken in isolation.

In compound concept detection area, some authors proposed a novel ontology-based Compound Concept Semantic Similarity calculation approach called CCSS which exploits concept constitution features [3]. In this method, the compound is decomposed into Subject headings and Auxiliary words, and the relationships between these two sets are used to measure the similarity. By experiment, the relationship between statistical models are explored.

Some likelihood intervals methods, likelihood ratio test, and χ_2, test for collocation theoretically was summarized by [5]. These methods were used for extracting collocations form a large scale corpus in Chinese language.

In image processing area, variance analysis is used for image segmentation, especially for thresholding. Some techniques such as Otsu method use this analysis for separating pixels image in two classes, black and white pixels. In [6], two-dimensional (2D) Otsu method was used to correct the Otsu threshold in segmenting images of low signal-to-noise ratio, this method completely removed the noise in renal biopsy samples images.

3 Method and Experiments

Figure 1 shows the general methodology for ontology learning process in pedagogical domain. Three classes corpus was analyzed to obtain important concepts. Afterwords, these concepts were joined to design an ontology, which is focused in the first step of the methodology: Corpus creation.

Corpus was created by a Web crawler to retrieve academic papers in Spanish. Those papers are focused in one of the following topics: Intelligence types, learning styles or teaching strategies. For obtaining principal concepts, a gold standard created by a domain expert was designed. Some of the principal concepts are composed by two or more words, then, a methodology for extracting automatically collocation was necessary.

Figure 2 shows the general design for automatic collocation detection process. The main idea is based on the selection of a threshold named α for dividing the words pairs in two classes: collocations and no collocations. Words are combined in pairs, and the threshold is obtained by analyzing the variance for quartiles. Variance is calculated based on frequency and conditional probability of word pairs.

Figure 3 shows experimental process for the experiments. It consists of two steps, each of them gets a threshold; first through the frequency and then through the conditional probability. Two steps were implemented due to there are three-words collocations, then, first a two-words collocations were obtained, afterwards, these collocations were analyzed with its neighbors.

The initial corpus is composed by 207 academic papers in Spanish about intelligences types, learning strategies and learning styles. This paper shows learning styles class results, with 83 papers.

After the method stops, the candidate list is evaluated using a gold standard (203 words) and the analysis of a domain expert. For preprocessing step, stop words were deleted and a lemmatizer was used for obtaining lemmas. Finally, texts were separated in bigrams.

According to Figure 3, in each of the iterations, the pairs of words were ordered by frequency or conditional probability variance. Afterward, the quartiles were obtained, finally , the general variance and per quartile were calculated and com-pared with the general variance. The hypothesis is that if a group of words has a low variance, these pair of words are not representative for the principal topic, and is not necessary consider them like a collocation.

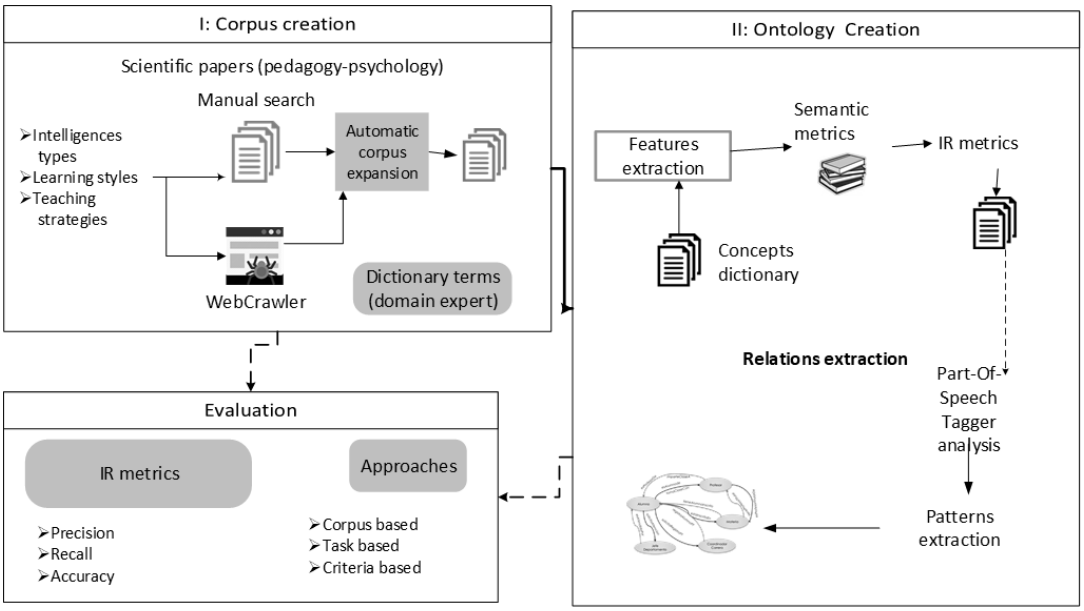

Fig. 1. Ontology learning methodology for pedagogical domain

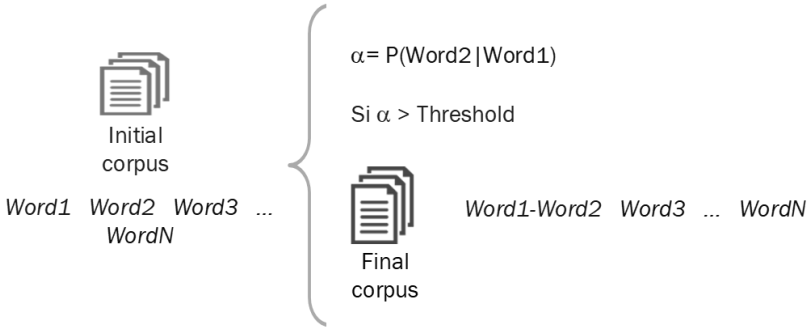

Fig. 2. General design for collocations detection

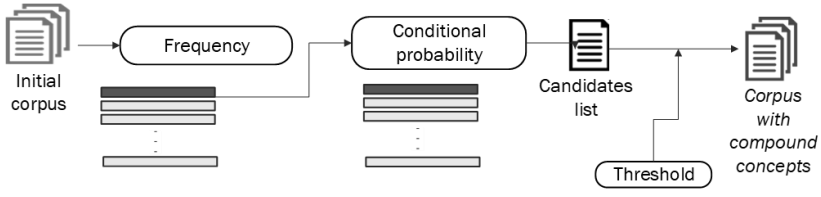

Fig. 3. Proposed method to detect collocations in a pedagogical domain corpus

ISSN 2007-9737

4 Results

In this section, a detailed discussion about results is presented. Table 1 shows data per quartile in first iteration, that is, number of quartile, minimum and maximum frequency, number of instances (the same for each quartile), frequency average and variance. Quartiles contain 24,022 pairs of words, most of them have a frequency of 1.

Table 1. Frequency results (first iteration)

Quartile	Minimum frequency	Maximum frequency	Frequency Average	Frequency Variance
1	2	3,054	4.729	22.8531
2	1	2	1.038	0.1912
3	1	1	1	0
4	1	1	1	0
Total	**1**	**3,054**	**1.9418**	**11.5397**

The greatest variation in frequency is in the first quartile, with a variance of 522 and an average of 5 times. According with the initial hypothesis, only the first quartile shows a frequency high enough for its word pairs to be considered as collocations, then, the remainder quartiles were deleted from the analysis.

The conditional probability of the first quartile 24,022 word pairs was calculated resulting in a behavior similar to that of the first iteration. That is, just the quartile 1 get a similar variance with the total result.

Table 2 shows these results. Bigrams of first quartile were analyzed respect to a threshold (α), which was established in the range $\chi - \sigma < \alpha > \chi + \sigma$. According with the data of first quartile, only the bigrams with a conditional probability between 0.105 and 0.7404 were retrieved as collocations.

Tables 3 and 4 show selected words pairs obtained in the experiments. Table 3 contains low conditional probability examples. Table 4 shows high conditional probability examples; and both of them have the method result (YES or NOT). 2,209 collocations were retrieved, the majority of them are domain independent, such as *property analysis*, *civil engineering* and *national university*.

Other collocations are principal concepts for the domain, and they are necessary to get an ontology, examples of them are *intuitive style*, *kinesthetic auditory*, *defined learning* and *elaborative thinking*.

Table 2. Obtained results for conditional probability (second iteration)

Quartile	Minimum probability	Maximum probability	Probability Average	Probability Variance
1	0.0952	1.0000	0.4211	0.3172
2	0.0228	0.0952	0.0474	0.0196
3	0.0071	0.0229	0.0134	0.0044
4	0.0003	0.0071	0.0034	0.0020
Total	**0.0003**	**1.0000**	**0.1213**	**0.2355**

Table 3. Examples of collocations in second iteration (low conditional probability)

Concept	Conditional probability	Result
Easier learning	0.0003	NOT
Constructivist learning	0.0003	NOT
Predominant learning	0.0003	NOT
Learning conclusion	0.0003	NOT
Learning happens	0.0003	NOT
Acceptable reliability	0.1905	YES
Interaction responds	0.1899	YES
Property analysis	0.1892	YES
Used simultaneously	0.1875	YES
Defined learning	0.1875	YES
Kinesthetic auditory	0.1875	YES
Civil engineering	0.7407	YES
Be advantageous	0.7368	YES
National University	0.7368	YES

Table 4. Examples of collocations in second iteration (high conditional probability).

Concept	Conditional probability	Result
Be able	0.7333	YES
Reflective active	0.7273	YES
Sleep quality	0.7273	YES
Elaborative thinking	0.7273	YES
Complutense University	0.7273	YES
Confirmatory factor	0.7222	YES
Intuitive style	0.1667	YES
Opposite style	0.1667	YES
Intuitive sensitive	0.1667	YES
Preferring face	0.1667	YES
Higher probability	0.1667	YES
Analyzed variables	1.0000	NOT
Inactive kinetic	1.0000	NOT
Level n31	1.0000	NOT

5 Conclusions and Future Work

In this paper, experiments to analyze the effect of variance metric in the automatic collocation detection was presented. The experiments were carried out considering papers in Spanish related to the learning styles. Finally, a list of collocation candidates was compared to a gold standard with the help of a domain expert.

The principal contribution of this paper is the use of variance analysis to obtain collocations automatically.

In both iterations, only a quartile gets a bigger variance respect to the average. Results contain principal concepts and independent domain collocations.

As future work, these experiments will be formalized in a methodology for collocations detention in other domains. In addition, experiments with more groups will be carried out, with the purpose to obtain an automatic threshold according with the domain behavior.

References

1. **Burger, W. & Burge, M. J. (2013).** *Principles of Digital Image Processing: Advanced Methods*. Springer Publishing Company, Incorporated.
2. **Dinu, A., Dinu, L., & Sorodoc, I. (2014).** Aggregation methods for efficient collocation detection. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 4041–4045.
3. **Li, M., Lang, B., & Wang, J. (2015).** Compound concept semantic similarity calculation based on ontology and concept constitution features. *International Conference on Tools for Artificial Intelligence (ICTAI)*, pp. 226–233.
4. **Pazos, J.-M. & Pamies, A. (2006).** Detección automatizada de colocaciones y otras unidades fraseológicas en un corpus electrónico. *Letras de Hoje*, Vol. 41.
5. **Yu, J., Jin, Z., & Wen, Z. (2003).** *Automatic Detection of Collocation*.
6. **Zhang, J. & Hu, J. (2008).** Image segmentation based on 2D Otsu method with histogram analysis. *CSSE (6)*, IEEE Computer Society, pp. 105–108. 978-0-7695-3336-0.

Article received on 29/10/2019; accepted on 07/03/2020.
Corresponding author is María Somodevilla.