International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:13, No:4, 2019

The Efficiency of Association Measures in Automatic Extraction of Collocations: Exclusivity and Frequency

Souhaila Messaoudi

Abstract**—**This paper deals with automatic extraction of 20 'adjective + noun' collocations using four different association measures: T-score, MI, Log Dice, and Log Likelihood with most emphasis on mainly Log Likelihood and Log Dice scores for which an argument for their suitability in this experiment is to be presented. The nodes of the chosen collocates are 20 adjectival false friends between English and French. The noun candidate to be chosen needs to occur with a threshold of top ten collocates in two lists in which the results are sorted by Log Likelihood and Log Dice. The fulfillment of this criterion will guarantee that the chosen candidates are both exclusive and significant noun collocates and thereby, they make perfect noun candidates for the nodes. The results of the top 10 collocates sorted by Log Dice and Log Likelihood are not to be filtered. Thereby technical terms, function words, and stop words are not to be removed for the purposes of the analysis. Out of 20 adjectives, 15 'adjective + noun' collocations have been extracted by the means of consensus of Log Likelihood and Log Dice scores on the top 10 noun collocates. The generated list of the automatic extracted 'adjective + noun' collocations will serve as the bulk of a translation test in which Algerian students of translation are asked to render these collocations into Arabic. The ultimate goal of this test is to test French influence as a Second Language on English as a Foreign Language in the Algerian context.

Keywords**—**Association measures, collocations, extraction false friends.

I. Introduction

THE collocations chosen to be translated in this test are 'adjective + noun' collocations, in which the adjectives are false friends with corresponding French adjectives.

Collocation as a notion has been defined in many different ways from different perspectives. In this study, the following definition will be adopted:

The tendency of a word be it lexical or functional to (significantly) re-appear in the company of another word within specific grammatical patterns within a specified proximity in a given corpus data. The word 'collocation', then, refers to any co-occurrence of words which is statistically (significantly) greater than would be predicted by their occurrence in all contexts, including all forms of formulaic expressions, multiword expressions (MWEs) and compositional phrases that may co-occur more than predicted even if they co-occur because of non-linguistic factors like real world facts (an example of collocation reflecting real-world facts being the co-occurrence of the verb sell and the noun house). This definition draws on existing definitions of 'collocation' in the literature, particularly [5], [6], [10]-[12], [14]. The reason behind choosing 'adjective + noun' collocations in this study is that this type of collocational pattern proved to be the most challenging collocation for native speakers of Arabic according to previous studies as noted in [1]-[3]. The 20 adjectival false friends chosen to be nodes in this study are part of a generated list of false friends between English and French. These false friends do not have any semantic or etymological overlap. Thereby, they are defined as strong false friends. The list of false friends consists of nearly 360 adjectives compiled in an Excel file format. Almost all these false friends are extracted from "Faux Amis and Key Words: A Dictionary-guide to French Life and Language through Lookalikes and Confusables" [15]. These adjectival false friends are classified into two types, weak and strong, based on whether they share some meaning/s or not, respectively. The 20 adjectival false friends chosen to be nodes in this study do not have any semantic or etymological overlap. They are, thereby strong false friends with corresponding adjectives in English. This research paper is an attempt to shed light on how these collocations have been constructed and the different steps applied to come up with a comprehensive list of sensible English collocations comprising these strong adjectival false friends.

S. Messaoudi is a Postgraduate Researcher at the Leeds University, Department of Arabic and Middle Eastern studies, School of Languages, Cultures, and Societies, United Kingdom (e-mail: mlsme@ university.ac.uk).

This study is part of a wider PhD project which is sponsored by a grant from the Algerian government and supervised by Dr. Claire Brierley and Prof. James Dickins.

II. Methodology

For the purposes of this study, Sketch Engine has been used for analyzing the data set using two large representative corpora for English and French languages respectively: English Web 2015 (enTenTen15) downloaded by Spider Ling in November and December 2015; and French Web 2012 (frTenTen12) crawled by Spider Ling in February 2012. Both corpora are encoded in UTF-8, cleaned, deduplicated, and tagged by Tree Tagger.

Most of the strong false friends' adjectives have been combined with their significant and exclusive noun collocates. The results, given by Sketch Engine for the potential collocates, drew heavily on the association measures available in this corpus tool.

Association measures are mathematical formulae in essence, which identify among different candidates in a corpus, the ones forming collocations through calculating

some scores expressing the likelihood of candidate phrases to be reliable collocations as mentioned in [8], [13]. This can be achieved through comparing the co-occurrence of the components of the candidate collocation with the co-occurrence of the two components together. Association measures can also be used for ranking data and classifying them. According to reference [13], the scores obtained by association measures indicate which collocations are at the top of the list. Association scores help as well with setting a threshold to discard all the collocations below.

A. Collocations Dimensions

1. Raw Frequency

The first is raw frequency which highlights the repetitive units generally in the corpus but may not be the best predictor of the regularity of occurrence and predictability in use, as in [7]. This can be shown through the fact that most occurrences of a particular collocation may occur in certain units in a particular text in a given corpus. Frequency, in such a case, cannot be a good indicator of collocability because the collocation in this instance is not equally dispersed across the whole corpus and is used only in one particular context. Although, frequency is one of the conditions in corpus linguistics to account for collocability, it is still affected by corpus size and how collocations are equally dispersed across the corpus [7]-[9].

2. Exclusivity

The second dimension for collocability is exclusivity. The idea of exclusivity stresses the positive relationship between the co-occurrence of two units in each other's company and their occurrence separately in the corpus [7]. This is referred to as "degree of exclusivity". It is also referred to as mutual information. Exclusivity is typically measured by the number of times each unit in a bigram appears on its own in the corpus compared with the number of their co-occurrences as defined in [7]. Therefore, the more two units co-occur, the stronger their exclusivity becomes.

3. Directionality

The third dimension is directionality. Directionality stems from the idea that components of collocations do not attract each other with equal strength; the degree of predictability within one collocation varies from one component to another. This can be explained through the strong predictability of one component in a collocation co-occurring with a specific word, while this specific word does not occur with very high frequency with the first component. An example that clarifies directionality is the word 'affair', which is highly likely to occur with 'love' more than the word 'love' does with affair as stated in [4, p. 141]. Therefore, 'affair' attracts 'love' more than 'love' does with 'affair'.

4. Compactness or Proximity

Compactness is the third dimension for collocability which is needed to identify collocations [7]. There are two approaches for determining the proximity of the collocations: the window approach and the n-gram approach.

a) The n-Gram Approach

The first approach, which is called the n-gram approach, highlights all the adjacent words as well as bi-grams as a special case. An n-gram is a string of adjacent words, in which (n) stands for the number of words starting from one word up to n (any) number of words.

b) The Window Approach

The second approach, also referred to as the window approach, selects adjacent words in a given span or window, from left and right of the node or the target word. This approach covers a wider range of pattern possibilities and allows more flexibility for different grammatical patterns to appear than the n-gram approach as noted in [7], [9]. Since the aim of this study is to look for adjacent attributive adjectives that often precede the noun, a windowing approach of one or two words following and two words preceding the node has been applied.

B. Association Measures

The extraction of collocation relies on considering the above dimensions. These dimensions can be depicted in a range of association measures. An association measure is defined as a mathematical formula which identifies among different candidates in a corpus the ones forming collocations through calculating some scores expressing the likelihood of candidate phrases to be reliable collocations [8], [13]. This can be achieved through comparing the co-occurrence of the components of the candidate collocation with the co-occurrence of the two components together. Association measures can also be used for ranking data and classifying them. According to reference [13], the scores obtained by association measures decide which collocations are at the top of the list. Association scores help as well with setting a threshold to discard all the combinations below it.

In theory, creating an association measure accounting for all the dimensions of collocability measures mentioned above, would result in a perfect association measure in extracting collocation. In practice, however, there is no association measure that covers all the dimensions discussed above due to the very different nature of some dimensions. For the purposes of this study, four association measures scores are going to be considered T-score, MI, Log likelihood, and Log Dice; with much emphasis on the scores of Log Dice and Log Likelihood.

1. T-Score

This mathematical metric tests the null hypothesis as to whether the co-occurrences of the collocation components are true or merely a chance. This measure takes into consideration the probability of the raw co-occurrence of the collocations units compared with the product of probability of the occurrence of each unit given that this data is normally distributed as noted in [9]. This assumption may not be suitable for linguistic data as words may not be equally dispersed across the whole corpus. Therefore, the scores of T-score are affected in a way or another by the raw frequency scores. It is also worth noting that the score is influenced by

the corpus size and the results cannot be comparable across other corpora as demonstrated in [7].

2. MI

This measure coined the title of the association of strength as it measures the dependence of the collocation units through knowing how much information each of the unit provides about the other; so, if one of them is entirely independent of the other one that will ultimately demonstrate no shared information between the two. The MI, therefore, is not a frequency-based measure like the T-score and for this reason, it favors low frequency collocations. These scores result in assigning high scores to rare exclusive combinations. Although mutual information is standardized and comparable across corpora, the scale with which it works does not set either a theoretical maximum or minimum value. This requires more attention when interpreting high scores as this does not mean necessarily that the collocation is stronger as mentioned in [7].

3. Log Likelihood

The Log Likelihood is a test of significance. This association measure provides two opposite hypotheses. The first hypothesis formalizes the independence of one component`s (word's) occurrences against the other component`s (word's) occurrences. W1 (word 1) does not necessarily co-occur with W2 (word 2). The second hypothesis stipulates that the co-occurrences of both words are dependent on each other. The Log likelihood score tells us how much more likely a collocation would occur under one of the hypothesis than the other, as stated in [9].

4. Log Dice

Log Dice is the measure of exclusivity. This explains why the mathematical expression of this measure looks similar to the one of MI. However, Log Dice came to compensate for the bias in favor of rare exclusivity which is the side effect of MI [7]. Moreover, when compared with T-score, and MI, Log Dice does not assume the random distribution model of the language that other measures invoke. This can be shown through the non-inclusion of the expected frequency in its equation. One of the other strengths of Log Dice is standardization and fixedness of its maximum value (14) as stated in [7]. Being standardized means that the results can be directly compared across corpora. The fixedness of the maximum value means that it operates on a specified scale, the highest score of which is 14. Log Dice has also been useful for translation pairs in machine translation, as stated in [16].

To determine the best noun collocates for the 20 strong adjectival false friends, the concordancer has been manipulated using the expert option of Sketch Engine to set specific parameters. This option allowed for making various specifications on the lemmas and part of speech filter. In this study, the lemma filter was set on both sides of the node with a window of three tokens. Therefore, after inserting the adjective (node) in the query box and setting specific parameters (lemma filter and part of speech filter), the concordance demonstrated many concordance lines in which the node co-occurred. To obtain the most significant and exclusive candidates for the focal adjective, the function of collocation that allows for choosing specific mathematical metrics of collocability, has been used (Log Dice and Log Likelihood).

For each node adjective, two lists of the top 10 noun collocates are generated. The results of the first list are sorted according to Log Likelihood scores. In the second list, however, the results are sorted by Log Dice scores.

The choice of the potential collocates in this experiment is just confined to nouns occurring in both lists.

TABLE I
LIST OF BOTH EXCLUSIVE AND FREQUENT NOUN COLLOCATES EXTRACTED BY MEANS OF CONSENSUS OF LOG LIKELIHOOD AND LOG DICE

N	Association Measures Collocations	T SCORE	MI	LOG LIKELIHOOD	LOG DICE
01	Heinous crime	88.227	12.374	121.960.295	7.422
02	Sensible	/	/	/	/
03	Actual Cost	88.995	5.642	48.433.387	6.087
04	Concurrent Enrollment	66.983	11.434	62.484.954	8.464
05	Candid camera	19.681	10.223	4.728.330	6.183
06	Comprehensive approach	137.653	6.976	148.509.366	7.326
07	Consistent manner	60.222	8.119	33.908.168	6.576
08	Eventual winner	57.106	8.987	34.373.167	6.380
09	Excited anticipation	19.354	10.792	4.866.526	6.313
10	Fastidious	/	/	/	/
11	Formidable Opponent	35.638	11.458	17.690.542	7.538
12	Rude awakening	49.293	14.641	44.810.151	9.648
13	Grand prize	132.041	10.282	215.334.403	9.035
14	Inhabited Island	2.445	9.395	66.267	6.144
15	Jolly	/	/	/	/
16	Rentable	/	/	/	/
17	Secular humanism	45.481	13.153	33.780.162	8.548
18	Sympathetic ear	37.348	9.110	14.915.143	6.139
19	Ulterior motive	75.497	16.426	125.291.875	10.080
20	Petulant	/	/	/	/

III. RESULTS

Using the concordancer of Sketch Engine to look for the best noun collocate for the adjective "heinous" within the top 10 collocates at a span of one word on both sides of the node,

(2 tokens around the node word), shows that the word "crime/ crimes" is the strongest collocate for "heinous". All the chosen association measures assigned the highest scores to "crime" as being the most frequent by the T-score, strongest by MI, most significant by Log Likelihood and most exclusive by Log Dice. In this case, it is legitimate to describe "crime" as a valuable noun collocate for the adjective "heinous". The noun crime was ranked first among the top 10 collocates, in either cases, when the results were sorted by Log Dice and when sorted by Log Likelihood. The following table summarizes the scores given to the chosen collocations highlighted by both Log Dice and Log Likelihood.

The results came up with 15 collocations (as shown in the Table I) that are highly likely to be a source of confusion for Algerian translation students and even English language learners. The results proved as well that unlike T-score and Log Likelihood, which are two association measures based to a greater or lesser extent on frequency, Mutual Information (MI and Log Dice) are exclusivity-based association measures to a lesser or greater extent at the expense of frequency. For this reason, highlighting collocates that are both exclusive and frequent looks to be hard, if ever possible.

In theory and practice, Log Dice tries to compensate for the bias of low frequency, which is the main side effect of MI, by highlighting exclusive collocates that are not necessarily rare ones. It is for this reason that Log Dice scores do not show as many technical terms as the MI scores do. Therefore, Log Dice can be regarded as the best association measure for exclusivity and this is why some differences can be highlighted regarding the performance of MI and Log Dice although they share a similar approach (exclusivity). However, Log Likelihood and T-score proved to have very similar results due to the very similar approach (frequency), even if the Log Likelihood scores have more credibility than those of the T-score. Log Likelihood, thereby looks to be a perfect measure among the associations of frequency.

With 20 strong false friends, 15 collocations have been formed based on the agreement of the Log Dice and Log Likelihood scores on the top 10 collocates for each adjective (as shown by Table I). Four instances, however, demonstrated non agreement about the top 10 collocates for four adjectives because each measure highlighted different collocates. Only one instance demonstrated that both Log Dice and Log Likelihood failed in giving a sensible list of collocates, either because the list was overwhelmed by function words by Log Likelihood or by technical terms by Log Dice. Although Log Dice and Log Likelihood do not agree either in principle or in approach, the results demonstrated that these two associations can agree in practice.

IV. Conclusion

Surprisingly, in most cases, comparing the results of Log Likelihood being one of the best measure of frequency, and Log Dice being one of the best measures for exclusivity, could agree about the most significant and exclusive collocations at the same time within the top 10 first collocates, in spite of their very different nature. The Log Likelihood frequency-based approach is demonstrated through returning function words. In the same vein, Log Dice measure, which is of an exclusivity nature, returns content words as opposed to Log Likelihood scores.

The results obtained from this study invite adopting a mixed approach of frequency and exclusivity together to yield more satisfactory results in extracting collocations.

References

[1] Ahmed, Z. A. A. 2012. *English lexical collocation knowledge of Libyan university students*. thesis, Prifysgol Bangor University.

[2] Al-Kattan, A. B. 2007. The Notion of Collocation in English with Reference to Arabic. *Buhuth Mustaqbaliya Scientific Periodical Journal,* 4(1), pp.7-17/18.

[3] Alsakran, R. A. 2011. The productive and receptive knowledge of collocations by advanced Arabic-speaking ESL/EFL learners. *Unpublished Thesis, Colorado State University, Colorado.*

[4] Brezina, V., McEnery, T. and Wattam, S. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), pp.139-173.

[5] Evert, S. 2008. Corpora and collocations. Corpus linguistics. An international handbook, 2, pp.1212-1248.

[6] Evert, S. and Kermes, H. 2003. Experiments on candidate data for collocation extraction. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2: Association for Computational Linguistics, pp.83-86.

[7] Gablasova, D., Brezina, V. and Mcenery, T. 2017. Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. Language Learning, 67(S1), pp.155-179.

[8] Gries, S. T. 2013. 50-something years of work on collocations. International Journal of Corpus Linguistics, 18(1), pp.137-166.

[9] Manning, C. D. and Schütze, H. 1999. Foundations of statistical natural language processing. MIT press.

[10] McEnery, T. and Hardie, A. 2012. Corpus linguistics: method, theory and practice. Cambridge: Cambridge University Press.

[11] Nesselhauf, N. 2005. Collocations in a learner corpus. John Benjamins Publishing.

[12] Pastor, G. C. 2017. Collocational Constructions in Translated Spanish: What Corpora Reveal. In: R. MITKOV, ed. Computational and corpus-based phraseology. London: Sprigner, pp.29-40.

[13] Pecina, P. 2010. Lexical association measures and collocation extraction. Language Resources and Evaluation, 44(1), pp.137-158.

[14] Sinclair, J. M. 1991. Corpus, concordance, collocation. Oxford: Oxford University Press.

[15] Thody, P., Evans, H. and Rees, G. 1985. Faux Amis and Key Words: A Dictionary-guide to French Life and Language Through Lookalikes and Confusables. Bloomsbury Publishing.

[16] Petrović, S., Šnajder, J. and Bašić, B. D. 2010. Extending lexical association measures for collocation extraction. Computer Speech & Language, 24(2), pp.383-394.