Comparative Evaluation and Integration of Collocation Extraction Metrics

Victor Zakharov(✉)

Saint-Petersburg State University, Universitetskaya emb., 7-9, 199034 Saint-Petersburg, Russia
v.zakharov@spbu.ru

Abstract. The paper deals with collocation extraction from corpus data. A whole number of formulae have been created to integrate different factors that determine the association between the collocation components. The experiments are described which objective was to study the method of collocation extraction based on the statistical association measures. The work is focused on bigram collocations. The obtained data on the measure precision allow to establish to some degree that some measures are more precise than others. No measure is ideal, which is why various options of their integration are desirable and useful. We propose a number of parameters that allow to rank collocates in an combined list, namely, an average rank, a normalized rank and an optimized rank.

Keywords: Collocation extraction · Association measures · Evaluation · Ranking · Average rank · Normalized rank · Optimized rank

1 Introduction

Let's speak about the notion of collocation. There are different approaches to this term. Sometimes a collocation is meant as a synonym of a word combination, sometimes it is a special type of a set phrase. S. Evert suggests the following definition: "A collocation is a word combination whose semantic and syntactic properties can't be fully predicted on the basis of information about its constituents and which therefore should be added to the dictionary (lexicon)" [1: 17]. But there are many set phrases whose meaning is equal to the sum of the meanings of their constituents, despite the fact that such phrases function as a single unit, with the stability rather than idiomatic nature being the main feature. A threshold of stability should be chosen to range them, above which a word combination can be called a set phrase. This approach assumes a probabilistic nature of collocations. Many modern authors and most of corpus linguists understand collocations as statistically determined set phrases. In this case, not only idioms but also multiword terms, named entities (real-world objects, such as persons, locations, organisations, products, etc.,) and other types of free combinations could be regarded as set phrases.

The above approach is the basic point of our paper which is aimed at evaluation of various statistical methods of automatic collocation extraction.

K. Ekštein and V. Matoušek (Eds.): TSD 2017, LNAI 10415, pp. 255–262, 2017.
DOI: 10.1007/978-3-319-64206-2_29

2 State of the Art

Nowadays, there are several ways to calculate the degree of coherence of parts of a collocation. A whole number of formulae have been created to integrate different factors that determine the association between the collocation components. Usually, such formulae are called association measures. P. Pecina provides 82 measures, and describes their mathematical foundations including their formulae and key references [2: 44–45, 48]. The most popular measures seem to be *MI, t-score,* and *log-likelihood*.

One should not forget also that words which tend to collocate with each other cannot be found in a random order in any case, as there exist grammar rules which imply that "the language system is a probabilistic one and it is a grammatical probability that word frequency shows in a text" [3: 31]. There are methods that take into account the syntactic nature of collocations. B. Daille claims that the linguistic knowledge drastically improves the quality of stochastic systems [4: 192]. One of the methods to take syntax in account are so-called word sketches, which are lists of statistical collocations, each one for each syntactic relation [5]. These syntax-based collocations are described in detail by V. Seretan [6: 59–101]. But in this paper, the grammatical probability is not taken into consideration, only the statistical one.

Lexical association measures being applied to a key word (node) occurrence and context statistics extracted from the corpus for all collocation candidates result in their association scores. But proper formulae are different, which is why collocation ranks obtained by different measures do not coincide. It is known, too, that some measures bring similar results and others are significantly different [7: 246–247].

The research on and evaluation of various association measures has been done for quite a long time and has been quite intensive. It is known that *t-score* extracts most frequent collocations. *Log-likelihood* was eventually preferred for its good behaviour on all corpus sizes and also for promoting less frequent candidates. On the contrary, the *MI* measure allows to reveal low-frequency multiword terms and proper names.

Besides, association score depends on the type of the units (lemmas or word forms) whose statistics are used for the calculations. The analysis described in [8: 340] has shown that in some cases word form collocations overwhelmingly have significantly bigger value.

The very number of the calculated collocates and the association scores are also dependent on the "window" between the node and the collocate that has been chosen for the calculations. When the window size is increased, besides meaningful syntagmas, words from a general lexico-semantic field are found as collocation candidates.

3 Collocation Extraction: An Experiment

The experiments were conducted on the basis of the Araneum corpora of Russian (http://unesco.uniba.sk), with the access provided through the NoSketch Engine [9]. We used 2 corpora, Russicum Minus (120 mln tokens) and Russicum Russicum Maius (1,20 bln). These corpora belong to the family of web corpora being created by the wacky technology [10].

Our objective was to study the method of collocation extraction based on the statistical association measures. We extracted collocations for the word *вода* (water) by means of the tool Collocations of the NoSketch Engine system using 7 association measures: *T-score, MI, MI3, log likelihood (LL), minimum sensitivity (MS), logDice and MI.log_f* [11].

The result for the query *вода* (water) was represented by a list of collocates (collocations) organized for each of the 7 above association measures ranged according to the association score in the form of a table (see an example in Table 1).

Table 1. List of collocates for *вода* (water) (a fragment)

Collocates	Co-occurrence count	Candidate count	MI.log_f score
Сточный (sewer)	12479	13791	100,505
Питьевой (drinkable)	11288	14006	97,878
Грунтовый (ground)	8672	11598	94,132
Кипяченый (boiled)	3635	4502	86,016
Горячий (hot)	20665	102240	84,393
…	…	…	…

A rank has been assigned to every collocate (i.e. collocation) according to the score of the each measure. The number of ranked collocates for each measure was 100.

4 Evaluation of the Effectiveness of Association Measures

Usually, comparison to some "gold standard" or expert evaluation are used to evaluate the results of automated systems. When methods of collocation extraction are evaluated both options appear to be problematic. There is no "gold standard" that would fully or significantly cover the set phrases. We could try to build it *ad hoc* for selected key words based on various dictionaries, but, due to the incomplete nature of dictionaries, the quality would be doubtful. As to expert evaluation, it is very expensive, taking into account time and human resources. Unfortunately, the quality of automated methods is often evaluated based on the examples taken from the top units of ranked lists, and from a small number of the resulting collocates [6: 70].

In this work, we have used expert evaluation on rather big amount of collocations obtained, namely, 100 for each measure. Further, we calculated the number of "true" collocations for each measure individually (Table 2).

The sum in each column can be interpreted as the precision indicator (in percentage) for the upper part of the ranked list.

However, it is not only the number of the true collocations extracted using each measure that is important: the rank of the relevant collocations is significant, too. This is why it would be prudent to introduce a weight of true collocations for each measure taking into account the place of the collocates in a sorted table. In order to evaluate the efficiency of each of the association measures the Kharin-Ashmanov method, which evaluates the relevance of the information retrieval results, was used [12].

Table 2. Distribution of the number of "true" collocations for each measure

Ranks	T-score	MI	MI3	LL	MS	log-Dice	MI.log_f
1–10	0	5	4	2	5	6	8
11–20	4	3	4	4	2	2	3
21–30	2	2	3	3	4	1	4
31–40	1	7	4	3	0	4	6
41–50	1	1	3	2	2	2	2
51–60	2	5	2	4	1	2	2
61–70	0	5	1	4	0	2	1
71–80	3	4	7	0	2	1	3
81–90	1	4	4	2	2	2	1
91–100	3	3	1	2	3	0	1
Total	17	39	33	26	21	22	31

Based on the expert evaluation of the extracted collocates and their place in the ranked list with regard to each association measure, a characteristic set was formed. A characteristic set for each measure means the number of the true collocations from the ranked list (precision value) obtained with this measure. According [12], we select characteristic sets that contain 5 elements – that is the precision values for the first 10, 30, 50, 70 and 100 collocates from the top of the list.

A weight is assigned to each element of the characteristic set (5, 4, 3, 2, and 1, respectively). Each element is "weighed": each of 5 precision values is multiplied by the its weight and divided by 15 (the sum of all weights). The sum of the weighed elements is the resulting precision of the characteristic set, i.e. the precision for appropriate measure.

Here is an example for the *MI* measure that has 5 true collocates in the top ten candidates (precision is 0.5), 10 true collocates (5 + 3 + 2, see Table 2) in the top thirty (precision is 0.33), 18(5 + 3+ 2 + 7 + 1) in the top fifty (0.36), 28 in the top seventy (0.40), and 39 in the top hundred (0.39). Then, each element was normalized (weighed) and the resulting precision will be equal to 0.5 * 5/15 + 0.33 * 4/15 + 0.36 * 3/15 + 0.4 * 2/15 + 0.39 * 1/15 = 0.167 + 0.088 + 0.072 + 0.053 + 0.026 = 0.406.

The values of the precision calculated like that for all seven measures are given below (Table 3).

Table 3. Precision values for association measures

	t-score	MI	MI3	LL	MS	log-Dice	MI.log_f
Precision	0.115	**0.406**	0.366	0.262	0.357	**0.391**	**0.562**
Place	7	2	4	6	5	3	1

So, in this case the best measures seem to be *MI.log_f, MI* and *log-Dice.* Of course, this result based on a single keyword is not enough for a safe generalization. Nevertheless, experiments with other keywords mostly confirm above list adding to it *min. sensitivity* (the sequence of measures can differ).

5 Integration of Different Association Measures

The next part of the study is aimed at developing methods for the integrated use of different measures of association. We used 7 collocation lists obtained in the first experiment. The lists of collocates were processed in the following manner. Meaningless collocations with punctuation marks were removed. Due to errors of lemmatization, some collocates were presented in several different word forms. For such cases, non-lemmatized word forms of the same word were united into a single unit, with the highest association value being chosen. "Clean" ranged lists of collocates were obtained as a result. Then, 7 tables (with 100 collocates in each) were merged into a new one in such a way so as to the collocates that were obtained through several measures were merged into a single line of the combined table, with their rank for each measure being provided. When a collocate was not available among the first hundred collocates for some measure it had no rank (see Table 4). The combined table counted 247 collocations. By the way, according to expert evaluation 86 of them were marked as true.

Table 4. Combined table of collocates for *вода (water)* with integrated ranks (a fragment)

Collocates	T-score	MI	MI3	LL	MS	log-Dice	MI.log_f	*n*	R_{av}	R_{norm}
Сточный (sewer)	5	25	1	2	5	4	1	7	6.14	6.14
Питьевой (drinkable)	7	39	2	4	7	6	2	7	9.57	9.57
Грунтовый (ground)	13	53	4	7	13	10	3	7	14.71	14.71
…	…	…	…	…	…	…	…	…	…	…
Родниковый (spring)	-	78	70	-	-	-	30	3	59.33	103.23
Туалетный (cologne)	73	-	37	45	75	57	31	6	53.00	59.36

It is clear that the same collocations with the word *вода* in the ranked lists of different measures have different rank, i.e. different measures estimate the syntagmatic association strength (collocability) between the components of a collocation in a different way. So, there is an idea that the collocation lists obtained through different measures should be merged. Then a question arises: what is the rank of a certain collocation in such merged list, or, in other words, what unique single rank should be assigned for each collocation.

The following hypotheses were made:

(1) the more the number of the measures that identified a relevant collocate, the stronger the collocability of a given collocation;
(2) the less the sum of the ranks or the average rank for a relevant collocate, the stronger the collocability;
(3) if both above conditions are observed then the "value" of a given collocation is higher, which is why we introduce the notion of a normalized rank.

As a result, the following indicators (parameters) have been added to combined table (Table 4):

(1) the number of association measures (*n*) that have "calculated" a given collocate (within first 100 lines);

(2) the *average rank* of the collocate (R_{av}): the sum of all non-empty ranks divided by their number;
(3) the *normalized rank* of the collocate.

The normalized rank (R_{norm}) is calculated as follows:

$$R_{norm} = k \cdot R_{av},$$

where k is the coefficient calculated by the following formula:

$$k = log_2(1 + 7/n),$$

where n is the number of the successful measures for this collocate.

It is safe to say that the average and the normalized ranks "objectify" (integrate) the functionality of various association measures.

6 Optimized Rank

However, ranks are based on a association measure score, which is why it is our task (including within this article) to correlate the ranks, i.e. the association strength, with some truth criterion concerning the effectiveness of appropriate measures.

The data on precision values for association measures (Sect. 4) allows to establish to some degree that such measures as *MI.log_f, log-Dice*, *MI*, and *MS* are more preferable. Having obtained "objective" evaluation of the efficiency of individual measures, we suggest introducing an indicator that is calculated taking into account the preference of the measures. We will call it the *optimized average rank*.

It is calculated as follows: all products of non-zero ranks multiplied by the coefficient of the measure significance are summed up and are divided into the number of measures used for a given collocate. In our case the measure significate coefficients are set, with their precision taken into account (Table 3): *MI.log_f* – 0.4, MI – 0.5, *logDice* – 0.6, *MI3* – 0.7, *min. sensitivity* – 0.8, *log-likelihood* – 0.9, *T-score* – 1.0. As a result, the rank of the collocations extracted by more efficient measures is reduced, and the relevant collocate in the combined table goes up. See the example in Table 5.

Table 5. Optimized rank for individual collocations

No.	Collocate	Average rank	Optimized rank
1.	Поверхностный (surface)	**81.5**	59.8
2.	Крещенский (baptismal)	82.0	**36.0**
3.	Обычный (usual)	**61.0**	34.1
4.	Газированный (sparkling)	63.0	**27.9**

Let's compare the collocates with even and odd numbers by pairs (*поверхностный vs. крещенский, обычный vs. газированный*). We can see that the latter, having collocations with *water* as the node, still have a bit higher average rank than the former. However, as per our suggestion, following the optimisation, the latter will have a lower

rank and go up in the ranked list (once again: the higher the rank in this list the higher the collocability degree). It seems the "odd" collocations really are stronger.

Naturally, it is so far only the idea. For real practice measure significate coefficients have to be chosen more reasonably, on a bigger experimental basis.

7 Conclusion and Further Work

To sum it up, the experiments have produced important results that characterise the efficiency of individual association measures. We also offer a method of assessing the effectiveness of statistical association measures.

Merging several lists of collocates obtained by different measures into one could improve the efficiency of statistical tools in total. We offer several options that allow to assess "the quality" of collocations in the combined list.

It is important to stress that the experiments were conducted using representative corpora, with large amount of the resulting collocations being under study. This was also confirmed in experiments with other words.

The evaluation procedure needs also special attention. Available lexical resources are both impure and incomplete. Therefore, the expert assessment remains one of the main methods but it needs thorough elaborated preprocessing and enrichment with terminological information.

Further research will be as follows:

1. Develop the programming tool that allows to make a single list of collocates with all the necessary parameters and to calculate integrated ranks.
2. Study how the efficiency of the association measures is associated with the width of the window (to the left and to the right of the key word) within which collocates are selected, and estimate the degree of such efficiency.
3. Identify the inter-relation between "syntagmatic" and "paradigmatic" collocates on the one hand and "idiomatic" and "statistical" on the other hand within the same search results, and identify the dependence of such inter-relation on the width of the window.

Acknowledgments. This work was partly supported by the grant of the Russian Foundation for Humanities (research project No. 16-04-12019).

References

1. Evert, S.: The statistics of word cooccurences word pairs and collocations. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung (IMS), Stuttgart (2004)
2. Pecina, P.: Lexical association measures and collocation extraction. Lang. Resour. Eval. **44**(1–2), 137–158 (2009). Prague
3. Halliday, M.: Current Ideas in Systemic Practice and Theory. Pinter, London (1991)
4. Daille, B.: Mixed approach for the automatic extraction of terminology: lexical statistics and linguistic filters [Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques]. Ph.D. thesis, Université Paris 7 (1994)

5. Kilgarriff, A., Tugwell, D.: Sketching words. In: Correard, M.H. (ed.) Lexicography and Natural Language Processing: A Festschrift in Honour of B.T.S. Atkins, pp. 125–137. Euralex, Goteborg (2002)
6. Seretan, V.: Syntax-Based Collocation Extraction. Text, Speech and Language. Springer, Dordrecht (2011)
7. Křen, M.: Collocation Measures and the Czech Language: Comparison on the Czech National Corpus data [Kolokační míry a čeština: srovnání na datech Českého národního korpusu], pp. 223–248. Kolokace, Praha (2006)
8. Zakharov, V., Khokhlova, M.: Syntagmatic relations in Russian corpora and dictionaries. In: Schoepe, K., et al. (eds.) Pragmantax II. The Present State of Linguistics and its Sub-Disciplines, pp. 333–344. Peter Lang, Frankfurt a.M. (2014)
9. Rychlý, P.: Manatee/Bonito – a modular corpus manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing, pp. 65–70. Masaryk University, Brno (2007)
10. Benko, V.: Aranea: yet another family of (comparable) web corpora. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2014. LNCS, vol. 8655, pp. 247–256. Springer, Cham (2014). doi:10.1007/978-3-319-10816-2_31
11. Statistics Used in Sketch Engine. https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/. Accessed 3 Feb 2017
12. Ashmanov, I., Grigoryev, S., Gusev, V., Kharin, N., Shabanov, V.: Using statistical method for intelligent computer-based text processing [Primenenie statisticheskih metodov dlja intellektual'noj komp'juternoj obrabotki tekstov]. In: The Proceedings of the Dialog 1997 International Seminar on Computational Linguistics and Its Applications, pp. 33–37 (1997)