Similarity between the Association Measures: a Case Study of Noun Phrases

Maria Khokhlova

Department of Mathematical Linguistics, St. Petersburg State University,
Universitetskaya emb. 11, 199034 St. Petersburg, Russia
m.khokhlova@spbu.ru

Abstract. Collocation extraction has gained much attention in natural language processing, its results are important in various areas of applied linguistics. The research focuses on a comparison between over a dozen of association measures based on a subset of the Russian Web corpus. The paper studies the automatically extracted Adj-Noun collocations. The aim of the experiments is two-fold. First, to examine the difference between statistical measures and second to find the most effective one for the Russian data. The former assumes the calculation of the Spearman's rank correlation coefficient and the latter implies the evaluation of the extracted lists against a Russian dictionary, i.e. identifying automatically extracted and manually collected collocations. The results are not such straightforward, one can distinguish between groups of measures that demonstrate a relative interchangeability. Also the produced bigrams can be considered as collocations by experts and thus may enrich dictionaries.

Keywords: collocability; collocations; corpora; statistics; statistical measures; gold standard

1 Introduction

Statistical tools play an active role in corpus linguistics and allow the researchers to extract data from texts supplying them with quantitative evaluation of the represented results. Collocation extraction is a task of natural language processing that is primarily based on statistical methods. Nowadays there are 82 statistical measures that are used for collocation extraction [1]. Usually they are called association measures and involve different principles. However, only a few of them were evaluated on linguistic data and even less applied to Russian. The aim of our experiments is to apply both well-known and not widespread association measures to the Russian data and to analyze possible similarity in the results.

The paper is structured as follows. In the next section we give a brief outline of the experiments. Then, we describe results of the analysis paying attention to the difference between the measures. Finally, the last section concludes the paper with discussion and gives suggestions for future work.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018, pp. 21–27, 2018.

2 Experiments

By a collocation we understand a recurrent word combination and analyze bigrams involving Adj-Noun model. The experiments were based on a subset of the Russian web corpus (ruWaC) [2] that comprises 9.5 mln tokens. At first we extracted over 200,000 Adj-Noun combinations from the corpus and then cleared the list leaving 197,343 bigrams. Among other phrases we deleted those with punctuation marks, other parts-of-speech (due to the errors in lemmatization and morphological annotation) etc. There was a number of Belorusian examples that were also excluded (only those that use letters lacking in the Russian alphabet, e.g. ў and і). The following noun phrases can be seen as the examples after processing the list: *gumanitarnaya aktsiya* "humanitarian action", *legendarny geroy* "legendary hero", *professional'naya konsul'tatsiya* "professional consultation", *natsional'naya osobennost'* "national feature", *prikladnaya sistema* "applied system" etc.

We tested the following thirteen association measures implemented in the UCS tool [3]: mutual information (MI), MI2, MI3, t-score, z-score, minimum sensitivity (MS), Dice, Jaccard and geometric mean (gmean) coefficients, Fisher, Poisson and chi-squared tests, and logarithmic odds ratio. The examined coefficients belong to different categories, e.g. exact and asymptotic hypothesis tests, point estimates of association strength, and heuristic measures. The comprehensive survey of the measures was made in [3,4]. To the best of our knowledge there is no comparison of these methods applied to the Russian language, however several measures were applied on Russian texts [5,6].

The experiments involved the comparison between each pair of measures in order to determine to what extent they produce the same results. Also we evaluated the extracted list across the dictionary [7] that can be seen to a certain degree as a source of true collocations. Dictionary data were automatically lemmatized by MyStem [8].

3 Results

3.1 Spearman's Rank Correlation Coefficient

We analyzed the lists of all bigrams extracted by the measures and calculated the Spearman's rank correlation coefficient (r_s) in order to assess the similarity between the measures. The coefficient can take values from -1 up to +1 indicating a perfect negative or a perfect positive correlation respectively. Zero value indicates there is no correlation between the ranks. Also we used co-occurrence frequency to demonstrate how the measures rank the extracted collocations compared to the simple metrics. The point of our work was also to study if the frequency can be applied instead of statistical measures or can be used as a baseline for further work on improvement of collocation extraction methods. The Table 1 presents the correlation between the pairs of measures.

Several pairs of the measures have the highest correlation that equals to 1 and thus show the same rankings. They are as follows: 1) Jaccard and Dice coefficients; 2) Poisson and Fisher tests; 3) chi-squared test, gmean coefficient,

Table 1: Values of Spearman correlation coefficient

	freq.	Dice	Fisher	Jaccard	MI	MI2	MI3	MS	Poisson	chi.sq	gmean	odds	t-score	z-score
freq.		0.21	0.56	0.21	-0.13	0.08	0.26	0.22	0.56	0.08	0.08	-0.13	0.77	0.08
Dice	0.21		0.74	1.00	0.72	0.77	0.79	0.99	0.74	0.77	0.77	0.64	0.60	0.77
Fisher	0.56	0.74		0.74	0.71	0.85	0.93	0.69	1.00	0.85	0.85	0.70	0.91	0.85
Jaccard	0.21	1.00	0.74		0.72	0.77	0.79	0.99	0.74	0.77	0.77	0.64	0.60	0.77
MI	-0.13	0.72	0.71	0.72		0.97	0.90	0.65	0.71	0.97	0.97	0.99	0.47	0.97
MI2	0.08	0.77	0.85	0.77	0.97		0.98	0.71	0.85	1.00	1.00	0.96	0.64	1.00
MI3	0.26	0.79	0.93	0.79	0.90	0.98		0.73	0.93	0.98	0.98	0.89	0.77	0.98
MS	0.22	0.99	0.69	0.99	0.65	0.71	0.73		0.69	0.71	0.71	0.57	0.57	0.71
Poisson	0.56	0.74	1.00	0.74	0.71	0.85	0.93	0.69		0.85	0.85	0.70	0.91	0.85
chi.sq	0.08	0.77	0.85	0.77	0.97	1.00	0.98	0.71	0.85		1.00	0.96	0.64	1.00
gmean	0.08	0.77	0.85	0.77	0.97	1.00	0.98	0.71	0.85	1.00		0.96	0.64	1.00
odds	-0.13	0.64	0.70	0.64	0.99	0.96	0.89	0.57	0.70	0.96	0.96		0.46	0.96
t-score	0.77	0.60	0.91	0.60	0.47	0.64	0.77	0.57	0.91	0.64	0.64	0.46		0.64
z-score	0.08	0.77	0.85	0.77	0.97	1.00	0.98	0.71	0.85	1.00	1.00	0.96	0.64	

z-score and MI2. Analyzing the data one can suggest that the coefficients within one group share some features and thus behaviour in common.

Jaccard and Dice coefficients are often viewed as similar statistics due to their formulae, the obtained results prove the fact showing that they are full equivalents when it comes to rankings. We find it peculiar that the value of r_s between Poisson and Fisher tests is so high (1.0).

In our experiments another four measures showed a strong correlation between them. The chi-squared test and z-score belong to asymptotic hypothesis tests, gmean coefficient exemplifies the degree of association group while MI2 is a pure heuristic statistic. Squares of z-score values equal to those of chi-squared and hence they rank collocations in the same way. As it was mentioned in [3] the gmean measure uses the geometric mean and is similar to MI. This statement holds true and here we find that the coefficient has even more in common with MI2.

As it was expected the co-occurrence frequency showed the lowest correlation with other measures with the exception of t-score. The t-test statistic can be seen as closely linked to the observed co-occurrence frequency (usually labelled as O11) and hence the high value of r_s (0.77) supports the statement indicating that the pair has strong positive correlation and produces similar ranking to a certain degree. The same negative value was obtained by the co-occurrence frequency in the pairs with MI and odds ratio (-0.13). That fact can suggest that the ranking produced by the measures do not coincide with the one made by co-occurrence frequency and moreover can be slightly the opposite. MI is often referred to as an association measure that is sensitive to low frequencies and

tends to overestimate them ranking on top rare word combinations. Also we find other values of r_s. The ones produces by the co-occurrence frequency with MI2, chi-squared test, gmean and z-score are extremely low (0.08) and can be interpreted as no correlation. Fisher and Poisson behave in between and indicate a middle correlation with the co-occurrence frequency.

MI achieved fairly strong correlation with almost all the measures. The largest values ranging from 0.90 up to 0.99 were shown by the pairs with MI2, MI3, chi-squared test, gmean coefficient, odds ratio and z-score. It is no wonder that MI correlates considerably with its heuristic variants (namely MI2 and MI3) that give greater weight in the numerator to O11. However r_s is lower for the pair MI and MI3 (0.90). One could not anticipate that r_s between MI and t-score will be relative high (0.47) as they are usually described as statistics placing opposite collocations on top.

Values of r_s for the pairs of t-score with other measures vary from 0.46 up to 0.91. This leads us to the conclusion that the coefficient behaves something in between but also provides the rankings that are similar to those produced by other statistics.

3.2 Top Bigrams Analysis

As next step we aimed to evaluate the extracted collocations ranked by different measures across the dictionary [7]. We analyzed true collocations (true positives) from top 100, i.e. collocations found in the dictionaries. The results showed the lists do not only contain true collocations but also meaningless combinations and the phrases that can be viewed as collocations but were not described in the dictionary. We put more emphasis on the third group expecting that such word combinations can be useful for lexicographic needs. We calculated true positives in top 100 across the dictionary and expert evaluation, Table 2 presents the results.

Co-occurrence frequency. The co-occurrence frequency produces positive results placing on top the following high frequent collocations that present in the dictionaries and are also marked by the expert: *tsennaya bumaga* "negotiable paper", *uchebnoye zavedeniye* "educational institution", *soyedinennye shtaty* "United States", *domennoye imya* "domain name", *meditsinskaya pomoshch* "medical care". Its precision for top 100 is high (0.75).

MI. MI proves the results found in other works retrieving the largest number of typos, mistakes in annotation and foreign lexis. Even though the Adj-Noun pairs were initially preprocessed the measure ranged top bigrams with Belorussian words and hapax legomena. The expert analysis marked the following collocations extracted by MI as true ones: *snezhnaya baba* "snowman", *Rizhsky balzam* "Riga balsam" (name of a Latvian herbal liqueur), *parnikovy gaz* "greenhouse gas", *finansovy defolt* "financial default", *pitschevye dobavki* "food additive". Also a large number of low-frequency proper names were found the list.

Table 2: Effectiveness of the association measures

Association measure	Precision (dictionary, top 100)	Precision (expert, top 100)	Precision (dictionary, all set)
frequency	**0.25**	0.75	**0.05**
Dice	0.00	0.28	0.01
Fisher	**0.25**	**0.87**	**0.05**
Jaccard	0.00	0.28	0.01
MI	0.00	0.30	0.01
MI2	0.00	0.28	0.01
MI3	0.22	0.82	0.03
MS	0.00	0.28	0.01
Poisson	**0.25**	0.86	**0.05**
chi.sq	0.00	0.28	0.01
gmean	0.00	0.28	0.01
odds	0.00	0.10	0.01
t-score	**0.25**	0.79	**0.05**
z-score	0.00	0.25	0.01

MI3. MI3 places on top more common collocations than two previously mentioned measures and extract noun compounds: *krugly stol* "round table", *mobil'my telefon* "mobile phone", *detsky sad* "kindergarden", *zapisnaya knizhka* "notebook", *zdravy smysl* "common sense". The precision of the coefficient is higher evaluated both against the dictionary and expert data.

MI2, MS, chi-squared test, gmean, Dice and Jaccard coefficients. As it was stated above the Dice and Jaccard coefficients give different values for bigrams but provide the same ranking due to their nature. They place on top combinations that are not listed in the dictionary and thus show low precision even against expert evaluation (0.28). The same score for precision was achieved by other statistics, as they produced the same lists being extremely sensitive to low frequencies of either nodes, collocates or their combinations.

Fisher and Poisson tests. The results show that the best precision was obtained by Poisson and Fisher coefficients and it holds true both for the dictionary and expert evaluation. Top 100 lists a high number of collocations, e.g. *krayn'aya mera* "extreme measure", *molodoy chelovek* "young man", *aktsionernoye obshchestvo* "joint-stock company", *postsovetskoye prostranstvo* "post-Soviet space", *tamozhennaya poshlina* "customs duty". It should be also noted that two measures rank collocations slightly differently but however 99% represent the same word combinations.

T-score. T-score has extracted the majority of high-frequent word combinations, e.g. *rossiyskaya federatsiya* "Russian Federation", *okruzhayutschaya sreda* "environment", *general'ny direktor* "general director", *vneshnyaya politika* "external politics", *intellektual'naya sobstvennost'* "intellectual property". The precision is also high and equals to 0.79.

Z-score. Being a close relative of t-score, z-score however does not yield the same results extracting several other collocations. The produced list of bigrams for top 100 coincides with the one by MI.

Logarithmic odds ratio. The coefficient overestimates low-frequency bigrams, however it also ranks top word combinations in which either a node or a collocate has extremely high values compared to other word. For example, we find *velikaya armada* "Great (Spanish) Armada", where the collocate "velikiy" (in masculine) has an absolute frequency 1,967 and the node "armada" only 1, hence the collocation occurs once.

3.3 Discussion

As we can see the precision of the obtained results is extremely low evaluated against the lexicographic data. This poor result can be explained by the structure of the dictionary entries that influenced the list of the collocations we used as a gold standard. The word combinations were automatically extracted from the dictionary part intended for idioms and phrases. Also compared to the corpus size the number of true collocations was quite low and it also made its impact.

According to the Zipf's law a vast number of lexis has low frequency and hence in a large corpus there is a certain number of words that occur only once. In case of association measures it can be the case that the 1st rank produced by a coefficient will correspond to several word combinations. The results reveal that MI2, MS, chi-squared test, gmean, Dice and Jaccard coefficients rank hapax legomena on top 100 and thus the bigrams coincide.

Values shown in the second and third columns correlate to a certain degree, i.e. for non-zero scores against the dictionary we find also better results given by the expert evaluation.

Top 100 bigrams extracted by the co-occurrence frequency, MI3, t-score, Fisher and Poisson coefficients proved to include much more true collocations than other measures. And that leads us to the conclusion that collocations described in dictionaries are frequent ones and thus they can be retrieved only by the measures that have strong correlation with frequency.

4 Conclusion and Future Work

In summary, the research shows that co-occurrence frequency, MI3, t-score, Fisher and Poisson tests yield significant collocates that occur relatively frequently. In

most cases, they are the most reliable measures. Our approach has its limitations as every dictionary is personalized and does not provide a comprehensive description of collocability. We believe it is important to study the coefficients on other large corpora and compare between them. Also corpus data should be cleaned up as the majority of measures in the experiments were sensitive to typos and errors. The size of the gold standard used for the evaluation should be increased, for now it is not sufficient enough.

The results demonstrate a relative interchangeability between the association measures and can be used in future work on quantitative methods and their evaluation. The possible solution for the improvement of collocation extraction techniques is to combine the measures, e.g. use more complex rankings involving values of different measures or add other models (syntactic or vector).

In future work we plan to make experiments with other languages and offer a wider scale comparison between them implying also more association measures and other types of phrases.

Statistical measures for evaluating the strength between the items can be used for word sense disambiguation, in translation studies and CAT systems, for identification of synonyms and antonyms etc.

Acknowledgements. This work was supported by the grant of the President of Russian Federation for state support of scholarly research by young scholars (Project No. MK-2513.2018.6).

References

1. Pecina P. Lexical Association Measures. Collocation Extraction. Prague: Institute of Formal and Applied Linguistics, 2009.
2. ruWaC: Russian web corpus, `https://www.sketchengine.eu/russian-web-corpus`
3. Evert S. The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, `http://purl.org/stefan.evert/PUB/Evert2004phd.pdf`
4. Evert S. Association measures, `http://collocations.de`
5. Khokhlova M. V. Eksperimental'naja proverka metodov vydelelnija kollokacij [Evaluation of Methods for Collocation Extraction]. In Slavica Helsingiensia 34. Instrumentarij rusistiki: Korpusnye podhody. Eds. A. Mustajoki, M.V. Kopotev, L.A.Birjulin, J.J. Protasova. Helsinki, 2008, pp. 343-357.
6. Pivovarova L., Kormacheva D., Kopotev M. Evaluation of collocation extraction methods for the Russian language. In Quantitative Approaches to the Russian Language (ed. by M. Kopotev, O. Lyashevskaya, A. Mustajoki). London, New York: Routledge, 2018. P. 137–157.
7. Yevgen'yeva A.P. (ed.-in-chief). Slovar' russkogo jazyka [A Dictionary of the Russian Language] vol. 1-4, 2nd edition, revised and supplemented. Moscow: Russkij jazyk, 1981-1984.
8. MyStem, `https://tech.yandex.ru/mystem/`