In Search of Lost Collocations: Combining Measures to Reach the Top Range

Eduard Klyshinsky
National Research University
Higher School of Economics
Tallinskaya str. 34,
123458 Moscow, Russia
+7(495) 7729590
klyshinsky@itas.miem.edu.ru

Maria Khokhlova
St. Petersburg State University
Universitetskaya nab., 11,
199034 St. Petersburg, Russia
+7 (812) 3289519
m.khokhlova@spbu.ru

ABSTRACT
The paper discusses statistical methods for collocation extraction. We test the following hypothesis: combining several methods gives a better result than applying just one. At the first stage we suggest two methods to combine MI and t-score rankings and evaluate the results on attributive and verbal collocations against the data attested in the dictionary. At the second stage, we use regression analysis to tune up coefficients that further improve the best method discovered at the first stage. These results are evaluated against native speakers' intuition and prove our main hypothesis for most cases.

CCS Concepts
• Applied Computing→Arts and Humanities • Applied Computing→ Document management and text processing.

Keywords
Collocations; statistical measures; evaluation; regression analysis.

1. INTRODUCTION
Since statistical metrics are a useful but not a precise tool for collocation extraction, there are a variety of association measures that are calculating collocation's words relatedness. As it was shown in [8], different indices are producing correlated collocation lists. Thus our intuition was that if we correctly combine several indices, it will increase overall precision of the collocation extraction.

Our study takes a look at two of the most frequently used methods, i.e., MI and t-score [2], and traces solutions for creating an aggregated list of collocations that best fits the gold standard. To prove the hypothesis, we offer two approaches that allow combining results, which we then evaluate against a dictionary and an evaluation of native speakers.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

DHS 2017, June 21-24, 2017, Saint Petersburg, Russian Federation
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5317-0/17/06…$15.00
DOI: 10.1145/3143699.3143713

2. COLLOCATION EXTRACTION METHODS
Numerous tools and methods have been used in automatic collocation extraction from corpora (e.g., see [12], for 47 statistical methods and [8], for 82 methods). In [8], the authors propose combining association measures for collocation extraction. They come to the conclusion that a neural network with five units in the hidden layer achieves the best result. Thus far, there has been no consensus on which method is the most suitable for this task. Depending on the data as well as the goal of a given research project, one or another measure may be better suited. For this research, we have taken only two measures—MI and t-score, which have been shown to have a minimal overlap in producing collocation lists. A t-score extracts the collocations used most frequently in a language. Its ranking is shown to be quite similar to the frequency ranking, although few differences in the rankings were crucial, since very frequent words, often matching a pattern by accident, are effectively filtered out. The MI measures the level of uncertainty in finding a collocate given a node. The MI refers to infrequent collocations and is highly sensitive to any noise; it should always be used with frequency filtering.

3. DATA
The experiments are based on the I-Ru corpus of approximately 156 mln. tokens morphologically annotated with TreeTagger [9]. For our purposes, we investigated token collocations of the selected nouns that have a meaning of body part, taking all verbal (V+N) and attributive (Adj+N) patterns into account, e.g., vz'erosit' volosy *'to tousle hair'*, židkaja boroda *'scraggly beard'*. Words denoting body parts tend to form many expressions; thus, we expected that this particular topic would yield sufficient data for our analysis. The following nouns were chosen: boroda *'beard'*, glaz *'eye'*, golos *'voice'*, krov' *'blood'*, lico *'face'*, noga *'leg'*, nos *'nose'* ruka *'hand'*, serdce *'heart'*, sleza *'tear'*, uxo *'ear'*, volosy *'hair'*, zub *'tooth'*.

In order to create aggregated collocation lists, we followed up the procedure as explained below. First, we extracted token collocates for the above-given nouns by applying MI and t-score to the data. Tokens of the same lemma were cut off except the one with the highest rank; tokens with Freq < 3 were completely excluded. The remaining lemmata in each list were ranked, and the top ten in each were taken for further investigation. In some cases, the highest available rank appeared to be less than ten, due to the lack of a given noun in the data (e.g., boroda *'beard'* has only four verbal collocates). With the raw frequency data subjoined for comparison, we obtained two ranked lists (MI and t-score), from

four to ten collocations in each, for 13 V+N and 13 Adj+N patterns.

4. RANK INDEX OF EVALUATION

Due to the different nature of produced by MI and t-score numbers, we cannot sum they up directly. That is why we have to use rank measures. To develop combined rankings, we used the rank sum (**Rsum**) and the minimal ranks (**MinR**), which are calculated as follows. Let RankMI(C) and RankT(C) be respectively ranks of collocation C according to MI and to t-score. The rank sum for a collocation C is the sum of its MI and t-score ranks: Rsum(C) = RankMI(C)+RankT(C). The MinR fixes the best position (i.e., a minimum rank number) for a given collocation in MI or t-score rankings. Let us suppose, for example, that a collocation has a higher rank by MI; then the t-score is seen as underestimating this collocation. The MinR is calculated as MinR(C) = min(RankMI(C), RankT(C)). Thus, for each noun a combined list of its collocates was calculated using Rsum and MinR; duplicates were cleaned up so that only higher ranked doubles remained on the list. These lists were compared with the selected dictionary in order to evaluate how full (recall) and precise (precision) they are[1].

Most of the previous evaluations of Russian data were based on the intuition of the evaluators or /and the available dictionaries (see [1]; [4]; [7]; [10]). In this part of research, we evaluate the performance vis-à-vis the collocation joint frequency, used here as a baseline, against "A Russian-English Collocational Dictionary of the Human Body" [3]. The given dictionary gives a good overview of the lexis referring to the human body and thus to the collocations. The results have been evaluated using two standard features: recall and precision. Recall is understood here as a number of collocations in the lists that are found in the Russian-English Collocational Dictionary. Table 1 demonstrates how much the results of the Rsum and minR methods, as well as the baseline, overlap with the Dictionary.

Table 1. Recall for Rsum and minR

Collocation Type	Baseline	Rsum	MinR
verbal collocations	0.46	0.47	0.43
attributive collocations	0.58	0.55	0.52

Precision takes into account the rank of a collocate in the lists of Rsum and MinR: the higher the value in the table, the higher the rank of the collocations attested in the Dictionary. For example, if collocations A, B, and C, attested in the Dictionary, take the first three places in the MI ranking, their Rsum is higher than for collocations X, Y, and Z, which occupy second, sixth, and tenth places in the MI ranking. These calculations are done using the mean reciprocal rank (MRR; see [11]), which is designed to calculate the probability of correctness, i.e. the overlapping of the ranked collocations with the Dictionary in each case. The rank is calculated as an average of the reciprocal ranks for the attested collocations; a higher value means that more attested collocations are aggregated at the top of the Rsum or MinR lists respectively. Table 2 shows the results of the proposed methods by comparison with the baseline.

Table 2. MRR for Rsum and MinR

Collocation Type	Baseline	Rsum	MinR
verbal collocation	0.15	0.17	0.16
attributive collocations	0.15	0.15	0.16

Both methods in total outperform the baseline (albeit not by much) in both recall and precision for verbal collocations, with the Rsum seeming to demonstrate better overall results. For attributive collocations, the results are more contradictory: the Rsum wins for the recall, while the best average precision is achieved by the MinR with both the Rsum and the baseline being equally behind. The collocations under consideration tend to have visible spreads in values; for example, verbal collocations of uxo *'ear'* or golos *'voice'* have the highest precision values when measured by the Rsum, while the verbal collocations of sleza *'tear'* demonstrate the highest values when the MinR is used. This also indicates the fact that there is no one best measure to extract collocations. A number of bigrams that are not described in the dictionary were also extracted by the measures. These phrases can also be taken as collocations and present in the top of the lists (this observation can partially explain low rate of precision).

5. REGRESSION ANALYSIS

5.1 Method

In this part, we aim to make the next step to our main hypothesis by calculating the regression coefficients for **Rsum**. The main idea behind this work is as following: Rsum produces results, which are still just a little bit better than the baseline. To further improve them, we choose coefficients for the MI and t-score rankings that allow picking up the best possible collocations from two lists. The formula is used in the following form: RRsum(C) = kMI * RankMI(C) + kT * RankT(C), where kMI and kT are regression coefficients (bearing 1 as a default value). We used MI and t-score values assigned to the collocations, which were summed up and applying coefficients in the experiment (Section 5.3). At the next stage we selected ten collocations high ranked by RRsum with the coefficients applied and evaluate these collocations against those, marked stable by the native speakers. The more coincidence we get the better results we achieve.

5.2 Evaluation by native speakers

In this part, we use evaluation against native speakers' in order to demonstrate their intuition about what are considered to be collocations. There might be many ways to understanding collocability in answering the question "Is bigram X a (lexical) collocation?" However, in a questionnaire, participants were asked to evaluate the given expressions and rate them on a scale from 1 to 5 that follows the classification developed by [6], except for linguistic jargon — a collocation was explained as a set of words that regularly co-occur regardless of underlying grounds (idiomatic or otherwise) for their co-occurrence. For the questionnaire, we used the data from [3]. In the survey, twenty automatically extracted collocates for each word and two

1 This dictionary was chosen because it is a practical realization of the Meaning-Text Theory, whereby lexical connections, or "lexical functions", are given undivided attention (see [6]).

distractors (added to control the quality of the output) were randomly presented for evaluation. Then a cross-agreement's value was calculated that means more confidence that a given combination is seen a collocation by the speakers' negative value means that a combination is not considered as a collocation at all. Since we do not split stimuli into any preset groups, we cannot break the inter-agreement values into corresponding classes; instead, we plot the response standard deviation against an average response value. The results of this are ranking lists of collocations that our respondents have marked as stable. In the following up evaluation these lists are used as a 'gold standard' in regression analysis presented below.

5.3 Results

Below we test our main hypothesis, that linear combination of two collocation indices produces better results, by varying the given coefficients kMI and kT. To do that a grid search was built over two-dimensional space in area [-10, 10] with step 0.1. On every step, the lists of collocations were ranked according to **RRSum**(C); native speakers' inter-agreement values from the Section 5.2 of the first ten collocates in the ranked lists were then summed up. The results for grid-search are demonstrating that results of MI and t-score are correlated, since visualization shows lines running from the center to the corners of the grid. This result coincides with conclusions made by other researchers (e.g. see [6]), who discovered that applying different measures gives in practice quite the same results.

However, a practical method has to define a formalism for selecting regression coefficients. Thus, we have evaluated the values at several points on this grid trying to find optimal values. Here and below we write a point as [kMI, kT]; e.g., [1.5, 1] means that kMI=1.5 and kT=1; [0, 1] means that kMI=0, and only t-score values are taken into account. When comparing to the native speaker's lists, we have found out that results are better in case of summing ranks for MI and t-score with coefficient that are both equal to 1. Therefore, for the given dataset optimal solution looks like 1.5 * x* RankMI(C) + x* RankT(C) where x is a real number (which is true at least for the interval [0, 20]).

Hypothesizing that the defined coefficients are the best solution for any subset, it means that for a whatever long list of collocations in a given corpus, we can generate a randomly selected subcorpus, tag it, and find an optimum in regression coefficients that can be applied for the list in whole. In order to prove this hypothesis we have generated several random sublists deleting 25%, 50%, and 75% of collocations from the used lists. The maximal value of the calculated sum was found by random search, since it is faster way that also keeps the comparable precision in our case.

Our hypothesis was proven for 4 of 5 (80 percent) points in the sublists with 25 and 50 percent data deleted, and the max value is reached twice for the sublist with 25 percent of the data deleted. That means that in most cases a set of coefficients [1.5, 1] with 1.5 for MI gives better results than using just raw MI and t-score measures separately or summing their ranks without coefficients. However, having 75% collocations deleted the coefficients follow to another best solution ([0, 1]) for three of five sublists. We performed the grid searches to illustrate these contradictory results. Continuum of maximal solutions is placed along [0, 1] axis, which means that t-score works fine without summing it up with MI. The key factor in such disappointing result is a well-known fact that less data drastically worsen results of a data-driven research.

5.4 Conclusion

MI and t-score enlarge the results found by each of the measures and therefore can be applied together to cover more collocations. Both methods of combining measures yield quite the same results, however they are better than the baseline. Regression analysis shows that the coefficients 1.5 and 1 for MI and t-score respectively give the best results on a big enough dataset. However, it is not always true. The longer list of collocations we have the better result we get in summing the lists extracted by two measures.

6. FUTURE WORK

We believe that the performance of a measure depends on the collocational preferences of a given token, i.e., on its general tendency to co-occur with other tokens. To exploit the difference between the distribution of features in the pattern vs. their distribution in the corpus as a whole, we used the Kullback-Leibler Divergence, whose reliability is proven in the morphological data in the work [5]. A future step in that direction would be to adopt the KLD values as a preprocessing coefficient that predicts whether or not a given token tends to form stable collocations. For example, the normalized KLD value for attributive tokens used with boroda *'beard'* is 2.68, while the normalized KLD for uxo *'ear'* is only 0.05. A lesser KLD means less confident results for whatever method is used to extract collocations, just because the noun itself tends to collocate to a lesser degree. The question is how to adapt the KLD to the collocation extraction methods discussed in this article.

Another methodological issue is a dataset available for evaluating purposes. Although any dictionary is a good example of expert knowledge in the field, it is generally acknowledged that this kind of source is often not comprehensive and has the disadvantages of being personalized and outdated after a lapse of time. Experiments with native speakers give insight into the current state of a language, but are much more difficult to conduct. The experiments have to be carefully planned, and there are always limitations on the number of examples that can be presented to the participants. A necessary methodological step that has to be taken is to create a gold standard that would work better in an evaluation process and avoid predictable, but irrelevant lower dictionary recall.

7. ACKNOWLEDGMENTS

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-5274.2016.6).

8. REFERENCES

[1] Braslavskij, P., and Sokolov, E. 2006. Comparing four methods for automatic extraction of two-word terms from a text [Sravnenie chetyreh metodov avtomaticheskogo izvlechenija dvuhslovnyh terminov iz teksta]. *Proceedings of the International Conference "Dialogue 2006"* [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2006"], 88–94.

[2] Evert, S., and Krenn, B. 2001. Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 188–195.

[3] Iordanskaia, L. N., Paperno, S., and MacKenzie, L. LaRocco, Leed J. 1996. *A Russian-English collocational dictionary of the human body*. Slavica Pub.

[4] Khokhlova, M. V. 2008. Evaluation of Methods for Collocation Extraction [Eksperimental'naja proverka metodov vydelelnija kollokacij]. In Slavica Helsingiensia 34. Instrumentarij rusistiki: Korpusnye podhody. Eds. A. Mustajoki, M.V. Kopotev, L.A.Birjulin, J.J. Protasova. Helsinki, 343–357.

[5] Kopotev, M., Pivovarova, L., Kochetkova, N., and Yangarber, R. 2013. Automatic detection of stable grammatical features in n-grams, *Papers from the 9th Workshop on Multiword Expressions (MWE 2013)*. Workshop at NAACL 2013 (Atlanta, Georgia, USA), June 13/14, 2013, Atlanta, 73-81.

[6] Mel'čuk, I. 1995. *The Russian Language in the Meaning-Text Perspective*. Wiener Slawistischer Almanach/Škola "Jazyki russkoj kul'tury": Vienna/Moscow, 682 p.

[7] Mitrofanova, O.A., Belik, V.V., and Kadina, V.V. 2008. Corpus analysis of selectional preferences of frequent words in Russian [Korpusnoe issledovanie sochetaemostnyh predpochtenij chastotnyh leksem russkogo jazyka]. *Proceedings of the International Conference "Dialogue 2008"* [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2008"], Vo. 7(14), 362–367.

[8] Pecina, P., and Schlesinger, P. 2006. Combining association measures for collocation extraction. *Proceedings of the COLING/ACL on Main conference poster sessions*, 651–658.

[9] Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. *Working Papers on the Web as Corpus*. Edited by Marco Baroni and Silvia Bernardini, 63–98.

[10] Toldova, S. Y., Akinina, Y. S., and Kuznetsov, I. O. 2013. The impact of syntactic structure on verb-noun collocation extraction. *Proceedings of the International Conference "Dialogue 2013"* [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], Vo. 12(19), 2–16.

[11] Voorhees, E.M. 1999. Trec-8 question answering track report. *Proceedings of the 8th Text Retrieval Conference*, 77–82.

[12] Wiechmann, D. 2008. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, 4(2), 253–290.