Various Criteria of Collocation Cohesion in Internet: Comparison of Resolving Power*

Igor A. Bolshakov[1], Elena I. Bolshakova[2],
Alexey P. Kotlyarov[1], and Alexander Gelbukh[2]

[1] Center for Computing Research (CIC)
National Polytechnic Institute (IPN), Mexico City, Mexico
{igor,gelbukh}@cic.ipn.mx
[2] Moscow State Lomonosov University
Faculty of Computational Mathematics and Cybernetics, Moscow, Russia
bolsh@cs.msu.su, koterpillar@gmail.com

Abstract. For extracting collocations from the Internet, it is necessary to numerically estimate the cohesion between potential collocates. Mutual Information cohesion measure (MI) based on numbers of collocate occurring closely together (N_{12}) and apart (N_1, N_2) is well known, but the Web page statistics deprives MI of its statistical validity. We propose a family of different measures that depend on N_1, N_2 and N_{12} in a similar monotonic way and possess the scalability feature of MI. We apply the new criteria for a collection of N_1, N_2, and N_{12} obtained from AltaVista for links between a few tens of English nouns and several hundreds of their modifiers taken from Oxford Collocations Dictionary. The nounits own adjective pairs are true collocations and their measure values form one distribution. The nounalien adjective pairs are false collocations and their measure values form another distribution. The discriminating threshold is searched for to minimize the sum of probabilities for errors of two possible types. The resolving power of a criterion is equal to the minimum of the sum. The best criterion delivering minimum minimorum is found.

1 Introduction

During the two recent decades, the vital role of collocationsin any their definitionwas fully acknowledged in NLP. Thus great effort was made to develop methods of collocation extraction from texts and text corpora. As pilot works we can mention [3,6,17,18]. However, up to date we have no large and humanly verified collocation databases for any language, including English. The only good exception is Oxford Collocations Dictionary for Students of English (OCDSE) [11], but even in its electronic version it is oriented to human use rather than to NLP. So the development of the methods of collocation extraction continues [4,5,9,12,13,14,15,16,19].

* Work done under partial support of Mexican Government (CONACyT, SNI, CGEPI-IPN) and Russian Foundation of Fundamental Research (grant 06-01-00571).

A. Gelbukh (Ed.): CICLing 2008, LNCS 4919, pp. 64–72, 2008.
© Springer-Verlag Berlin Heidelberg 2008

The well-known numerical measure of collocate cohesion used to extract collocations from text corpora is Mutual Information [10]. It is based on the ratio $(S \cdot N_{12})/(N_1 \cdot N_2)$ that includes numbers of collocates occurring closely together (N_{12}) and apart (N_1 and N_2), as well as the corpus size S.

However, the corpora, even the largest ones, suffer from data scarceness. Meanwhile, Internet search engines are considered more and more frequently as a practically unlimited source of collocations [7,8]. The transition to the Web as a huge corpus forces to revise all statistical criteria, since only numbers of relevant Web pages can be obtained from the search engines. The same words entering a page are indistinguishable in the page statistics, being counted only once, and the same page is counted repeatedly for each word included. Hence, Mutual Information measure is deprived of its statistical status. Therefore it is worthwhile to consider other cohesion measures (hereafter, we name them merely criteria) that depend on N_1, N_2, and N_{12} – now measured in pages – in a similar monotonic manner and retain so-called scalability feature of MI. Scalability is preserving the numeric value of a function with proportional changes of all its numeric arguments. This feature is required to diminish influence of systematic and stochastic variations of Internet statistics, since in each search engine the numbers N_1, N_2, and N_{12} for already well-known words are growing nearly proportionally over time.

The criteria to be chosen should have the most possible resolving power. It means that they should distinguish in a better way whether a given collocate pair is a true collocation or merely a pair casually occurred together. We could estimate the resolving power by the sum of probabilities for errors of the following two types: when a criterion considers a true collocation as false or when it considers a false collocation as true. So our plan is as follows.

We select a family of plausible criteria and prove that they possess the scalability and monotony against N_1, N_2, and N_{12}. Then we get a large set of triples N_1, N_2, N_{12} from AltaVista for collocate pairs formed by 32 English nouns and 1964 modifiers (mainly adjectives) that are recorded for these nouns in OCDSE. We consider the pairs that link the nouns with their own modifiers as true collocations, while 'noun–an alien modifier' pairs are considered false collocations. Some modifiers are common for several nouns, thus introducing errors in the attribution of some pairs. However, we neglect these facts since they affect all the criteria in a similar way.

In our experiments, the criterions values for 'noun–an alien modifier' pairs form one distribution, while 'noun–its own modifier' pairs form another. For the true pairs, any criterion usually gives greater values. A threshold is searched that minimizes the sum of probabilities for errors of the two types: attributing a false collocate pair to true collocations or a true collocate pair to false collocations. Resolving power of a criterion is defined to be that minimum. The best criterion delivers minimum minimorum.

It is shown that the best criterion unites N_1 and N_2 in the so-called harmonic mean. However, the remaining criteria under comparison give rather close results.

2 Various Numerical Criteria of Word Cohesion

Let us take the words W_1 and W_2 in a text corpus as would-be collocates and consider their occurrences and co-occurrences at a short distance as random events. Then their co-occurrence should be considered significant if the relative frequency N_{12}/S (= empirical probability) of the co-occurrence is greater than the product of relative frequencies N_1/S and N_2/S for the collocates taken apart (S is the corpus size in words). Using logarithm, we have the criterion of word cohesion known as Mutual Information [10]:

$$MI_{12} = \log \frac{S \cdot N_{12}}{N_1 \cdot N_2} \tag{1}$$

MI has an important feature of scalability: if all its building blocks S, N_1, N_2, and N_{12} are multiplied by the same positive factor, MI retains its value.

In the Internet we cannot evaluate events directly by numbers of words, since only Web page counts are available. Of course, we can re-conceptualize MI with all N being counts of the pages with relevant words or word combination and with S as the amount of pages indexed by the search engine. However, now N/S is not the empirical probabilities of word occurrence. We only cherish the hope that the ratio N/S is monotonically connected with the corresponding empirical probability for word occurrence.

An additional headache with MI is the page total S. Its evaluation is a separate task, necessitating several Internet queries. The substitution of S by the number of pages for the most frequent word in the given language (it is always an auxiliary word) does help [2], but the immanent Internet trend of volume growth keeps this additional measurement necessary. In such a situation, we are free to consider several different criteria built from the same numbers except of S, which we strive to exclude from the game. The sought-for criteria should:

1. Depend only on N_1, N_2, and N_{12};
2. Depend on N_{12} in a monotonously increasing manner;
3. Depend on N_1 and N_2 in a monotonously decreasing manner;
4. Depend on N_1 and N_2 in the same way, since we have no reason to consider any collocate more influential;
5. Be scalable.

So we change the ratio under the logarithm in (1) into the ratio N_{12}/M_{12}, where M_{12} is a specific mean value for the N_1 and N_2:

$$M_{12} = F^{-1}\left(\frac{F(N_1) + F(N_2)}{2}\right) \tag{2}$$

In (2), $F()$ is a monotonous function, and $F^{-1}()$ is its inverse. The features 1, 2, and 4 are evidently satisfied. The monotonous increment of M_{12} with growth of N_1 or N_2 (feature 3) can be shown through differentiating M_{12} by its arguments N_1 or N_2. It is interesting that the increment is valid even for any monotonously decreasing $F()$.

Table 1. Various types of the mean value

$F(z)$	M_{12}	**Name of** M_{12}
$\log z$	$\sqrt{N_1 N_2}$	Mean geometric
$1/z$	$2N_1N_2/(N_1+N_2)$	Mean harmonic
$\sqrt{z}$	$((\sqrt{N}_1+\sqrt{N}_2)/2)^2$	Mean square root
z	$(N_1+N_2)/2$	Mean arithmetic
z^2	$\sqrt{((N_1^2+N_2^2)/2)}$	Mean quadratic

However, the feature of scalability is not immanent for all types of $F()$, so we take only the specific group: $F(x) = \log x$ or $F(x) = x^p$, where p is positive. For them the scalability can be proved easily. Within the selected group, M_{12} coincides with well-known mean values (cf. Table 1). When collocates occur only together, so that $N_1 = N_2 = N_{12}$, the ratio N_{12}/M_{12} for all $F()$ in the group is equal to its maximum value 1. If these words never meet each other as close neighbors ($N_{12} = 0$), the ratio reaches its minimum value 0. When both words occur with nearly the same frequency, N_{12}/M_{12} is equal to N_{12}/N_1, which is usually a very small quantity.

To investigate the statistics of N_{12}/M_{12} in a more convenient way, we select logarithmic scale for it, just as for MI in (1), with the logarithmic base equal to 2 and an additive constant 16. Thus the collocation cohesion measure takes the form

$$CC = 16 + \log_2 \frac{N_{12}}{M_{12}} \tag{3}$$

The M_{12} in (3) is taken from Table 1, where the third column contains the name of the corresponding criterion.

The transformations in (3) put the maximum value to 16, while zero on the scale now corresponds to $N_{12} \approx N_1/65000$ in the case of $N_1 \approx N_2$. Previous research [1,2] of the geometric criterion with rather vast Web statistics gives evidence that the overwhelming majority of CC values for true collocations are in the interval $(0 \ldots 16)$. The minimal CC value goes to $-\infty$ because of the logarithm, so we may formally replace it by a large negative constant. We take -16, since this value was never reached for any positive N_{12} in our previous experiments.

It should be emphasized that all these scaling tricks in no way affect the further results. They merely expand the relevant scale interval and thus make it convenient for visual representation.

3 Modifier Sets Taken for Evaluations

We take as collocate pairs English nouns with their modifiers – both adjectives and nouns in attributive use – from OCDSE. The nouns were picked up in a rather arbitrary manner, with preference to those with larger modifier sets (cf. Table 2). The convenience of modifiers is that in English they frequently come just before its noun in texts, thus forming bigrams. A deeper research for

Table 2. Selected nouns and sizes of their modifier sets

SN	Noun	MSet Size	SN	Noun	MSet Size
1	answer	44	17	effect	105
2	chance	43	18	enquiries	45
3	change	71	19	evidence	66
4	charge	48	20	example	52
5	comment	39	21	exercises	80
6	concept	45	22	expansion	44
7	conditions	49	23	experience	53
8	conversation	52	24	explanation	59
9	copy	61	25	expression	115
10	decision	40	26	eyes	119
11	demands	98	27	face	96
12	difference	53	28	facility	89
13	disease	39	29	fashion	61
14	distribution	58	30	feature	51
15	duty	48	31	flat	48
16	economy	42	32	flavor	50

distant modifier pairs and collocations of other types necessitates considering word interval between collocates, and this essentially tangles the problem of evaluations of collocate co-occurrences through Internet search engines [2]. For these 32 nouns, total amount of modifiers, including repeated ones, is 1964 (1302 without repetitions). The mean modifier group size equals 61.4, varying from 39 (for *comment* and *disease*) to 119 (for *eyes*). The second and the third ranks determined by the set sizes correspond to *expression* (115) and *effect* (105).

Some nouns (*conditions*, *demands*, *enquiries*, *exercises*, and *eyes*) were taken in plural form in the experiments, since they are used with the recorded modifier sets in plural more frequently than in singular.

We have limited the number of nouns to 32 units, since the total amount of queries to the Web grows approximately as a square of this number. Taking into account the well-known limitations of Internet search engines, on the one hand, and the general trend of statistics growth, on the other hand, we have coped with ca. 50,000 accesses to AltaVista within a week, but we could not afford a greater task.

4 On Calculation of Resolving Powers

Our method of evaluation of the resolving power for various criteria is as follows. Let n_i, $i = 1 \ldots 32$, be nouns under research, and $M_{own}(n_i)$ be the sets of its own modifiers m_p. The set $M_{alien}(n_i)$ of modifiers m_q that are alien to n_i can be expressed by the formula

$$M_{alien}(n_i) = \bigcup_{j=1\ldots32, j\neq i} M_{own}(n_j)$$

We consider our five criteria, performing the following steps for each of them:

1. Calculate CC values for all pairs (n_i, m_p) and all i, forming the first distribution D_1.
2. Calculate CC values for all pairs (n_i, m_q) and all i, forming the second distribution D_2. It frequently contains the value $-\infty$ that corresponds to the collocate pairs never meeting together in the Internet closely (zero N_{12} value).
3. Changing threshold T by small steps, calculate the probability P_1 of D_1 tail in the region lower that T (this is the error of the first type, attributing a true collocate pair to false collocations), and the probability P_2 of D_2 tail in the region greater that T (this is the error of the second type, attributing a false collocate pair to true collocations) – cf. Figure 1. The minimal value of the sum $P_1 + P_2$ is the resolving power RP of the given criteria.

The RP values are then compared to each other and the minimum minimorum found, thus delivering the best criterion (champion). Note that $M_{alien}(n_i)$ can include some members of $M_{own}(n_i)$. The intersection of the sets increases the overlay of the distributions, but it does not eliminate their difference. Since the overlays affect the criteria in the same manner, they cannot change the champion.

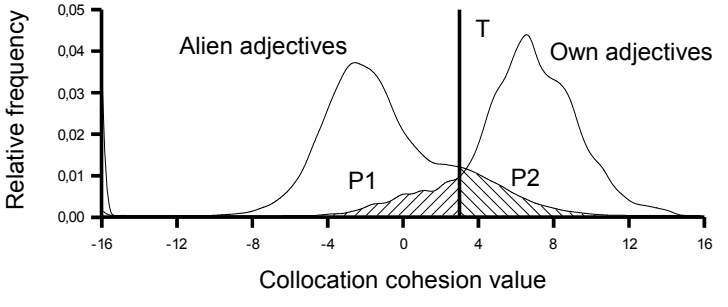

Fig. 1. Two distributions and the threshold

5 Experiment and Discussion of Various Criteria

The results of our calculation are given on Table 3. The best resolving power is delivered by the harmonic criterion with RP equal to 0.25 around the threshold 3.5. The worst is the quadratic criterion with RP equal to 0.30. We can see that the champion seems rather good, but the losers are not so far after. Moreover, shifts of thresholds in the intervals ± 2 centered at the minimums do not change the RP values significantly. All this means that collocation extraction from the Internet may be performed by any of these criteria with comparable results.

Our calculations also show that if CC for the champion is greater than 9.5, this pair is an obviously true collocation; and if it is lower than -3.5, the pair is an obviously false collocation.

Table 3. Resolving power of various criteria

Criterion Name	$F(z)$	Threshold for Minimum	Resolving Power
Geometric	$logz$	3.0	0.27
Harmonic	$1/z$	**3.5**	**0.25**
Square-root	$\sqrt{z}$	1.0	0.28
Arithmetic	z	1.5	0.29
Quadratic	z^2	1.5	0.30

The best criteria can be represented as (4)

$$CC = 16 + \log\left(\frac{N_{12}}{N_1 N_2}\frac{N_1 + N_2}{2}\right) \tag{4}$$

The comparison of (4) with (1) shows that the champion merely takes $2^{15}(N_1 + N_2)$ instead of S, with re-conceptualization of all numbers as measured in Web pages.

It is remarkable that the threshold 3.5 determined for the champion proved to be highly close to the threshold obtained in [2] for distinguishing true collocations from corresponding malapropos collocate pairs. To give a tip on the problem, let us consider a text with the malapropos phrase *travel about the* ***word***, where the intended ***world*** is erroneously replaced by the similar (paronymous) word ***word***. It is necessary to detect the pair *travel ...* ***word*** as false collocation and to propose the true collocation *travel ...* ***world*** as its correction. The detection of malapropos pairs and the search of their possible corrections can be done by means of cohesion measurement in the Internet, and appropriate experiments were carried out with representative sets of Russian malapropisms and with the aid of Yandex search engine.

Therefore, in [2] the close value of the threshold has been obtained for the definition of false collocations as malapropos pairs, for the different natural language, and for the different criterion (namely, the geometric one, cf. Table 3). This proves that the results of distinguishing correct collocations depend on natural language or criterion rather weakly.

6 Conclusions

We have proposed a family of numerical criteria to measure cohesion between words encountered in the Internet. All five criteria depend only on number of Web pages containing would-be collocates. The thresholds are found that minimize the sum of probabilities of errors of the two following types: considering a true collocation as false or considering a false collocation as true. The minimum is called resolving power RP of the given criteria. The best criterion delivers minimal RP among the peers. Its formula includes so-called harmonic mean for numbers of pages with collocate occurrences considered together or apart.

However, the remaining four criteria give comparable results. Therefore, each criterion among the considered ones may be taken for collocation extraction

from the Internet with nearly the same results. Further search of better criteria seems ineffective. The proposed criteria are applicable to different problems of computational linguistics, among them malapropism detection and computer-aided acquisition of collocations from the Internet.

References

1. Bolshakov, I.A., Bolshakova, E.I.: Measurements of Lexico-Syntactic Cohesion by means of Internet. In: Gelbukh, A., de Albornoz, Á., Terashima-Marín, H. (eds.) MICAI 2005. LNCS (LNAI), vol. 3789, pp. 790–799. Springer, Heidelberg (2005)
2. Bolshakova, E.I., Bolshakov, I.A., Kotlyarov, A.P.: Experiments in Detection and Correction of Russian Malapropisms by means of the Web. International Journal on Information Theories & Applications 12(2), 141–149 (2005)
3. Church, K., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1), 22–29 (1990)
4. Evert, S., Krenn, B.: Methods for the qualitative evaluation of lexical association measures. In: Proc. 39th Meeting of the ACL 2001, pp. 188–195 (2001)
5. Wu, H., Zhou, M.: Synonymous Collocation Extraction Using Translation Information, `http://acl.ldc.upenn.edu/P/P03/P03-1016.pdf`
6. Ikehara, S., Shirai, S., Uchino, H.: A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. In: Proc. COLING 1996 Conference, pp. 574–579 (1996)
7. Keller, F., Lapata, M.: Using the Web to Obtain Frequencies for Unseen Bigram. Computational linguistics 29(3), 459–484 (2003)
8. Kilgarriff, A., Grefenstette, G.: Introduction to the Special Issue on the Web as Corpus. Computational linguistics 29(3), 333–347 (2003)
9. Krenn, B., Evert, S.: Can we do better than frequency? A case study on extracting pp-verb collocations. In: Proc. ACL Workshop on Collocations (2001)
10. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
11. Oxford Collocations Dictionary for Students of English. Oxford University Press (2003)
12. Pearce, D.: Synonymy in collocation extraction. In: Proc. Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations. NAACL 2001, Pittsburgh, PA (2001), `http://citeseer.ist.psu.edu/pearce01synonymy.html`
13. Xu, R., Lu, Q.: Improving collocation extraction by using syntactic patterns. In: Proc. IEEE Int. Conf. Natural Language Processing and Knowledge Engineering, IEEE NLP-KE apos.05, pp. 52–57 (2005)
14. Seretan, V., Wehrli, E.: Accurate collocation extraction using a multilingual parser. In: Proc. 21st Int. Conf. Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, Australia, pp. 953–960 (2006)
15. Seretan, V., Wehrli, E.: Multilingual collocation extraction: Issues and solutions. In: Proc. Workshop on Multilingual Language Resources and Interoperability, Sydney, Australia, pp. 40–49 (2006)
16. Seretan, V., Nerima, L., Wehrli, E.: A tool for multi-word collocation extraction and visualization in multilingual corpora. In: Proc. 11th EURALEX International Congress EURALEX 2004, Lorient, France, pp. 755–766 (2004)

17. Smadja, F.: Retreiving Collocations from text: Xtract. Computational Linguistics 19(1), 143–177 (1990)
18. Smadja, F.A., McKeown, K.R.: Automatically extracting and representing collocations for language generation. In: Proc. 28th Meeting of the ACL, pp. 252–259 (1990)
19. Wermter, J., Hahn, U.: Collocation Extraction Based on Modifiability Statistics. In: Proc. 20th Int. Conf. Computational Linguistics COLING 2004, pp. 980–986 (2004)