Criteria of Word Connectivity Evaluations by means of the Web*

Igor A. Bolshakov and Vicente Cubells

Center for Computing Research (CIC)
National Polytechnic Institute (IPN), Mexico City, Mexico
igor@cic.ipn.mx, vcubells@yahoo.com

Abstract. Syntactic links between content words in sensible texts are intuitively conceived 'normal,' thus ensuring text connectivity. Criteria of the word connectivity are already known for text corpora, but for the Web they should be revised. We propose to measure lexico-syntactic connectivity between content words by means of the Web with a specially introduced Stable Connection Index (*SCI*). *SCI* is similar to Mutual Information well-known in statistics, but does not require iterative evaluation of total amount of Web-pages under search engine's control and is insensitive to both fluctuations and slow growth of raw Web statistics. Based on Russian material, *SCI* exhibited concentrated bell-form distribution for some type of word combinations. We compare *SCI* criterion with other one which, being very similar to Mutual Information in form, proved to be nearly equivalent to *SCI* in its distribution.

1. Introduction

Syntactic links between content words in sensible texts are intuitively conceived 'normal,' thus ensuring text connectivity. In fact, the task of measurement of connectivity between words arose many years ago, in information retrieval (e.g., [3, 10]) and, in the recent decade, in collocations extraction and acquisition (e.g., [4, 6, 9]). For text corpora evaluations, a well-known statistic measure reckoning on numbers of occurrences of words and their combinations was proposed under name of Mutual Information [9]. Rapid development of the Web compels to revise the available methods and criteria oriented to the text corpora [5]. Indeed, raw statistical data in Web search engines are measured in numbers of relevant pages rather then in word occurrences.

In this paper we propose to measure lexico-syntactic connectivity between content words with a specially introduced Stable Connection Index (*SCI*). In the form, *SCI* is similar to Mutual Information (*MI*), but it operates by statistics of relevant Web-pages and thus is convenient for Web measurements. Just as *MI*, *SCI* is nearly insensitive to both quick fluctuations and slow growth of all raw Web statistics delivered by a search engine, but in contrast to *MI*, it does not require iterative evaluation of the total amount of pages under the engine's control. Our additional goal is to compare *SCI*

* Work done under partial support of Mexican Government (CONACyT, SNI) and CGEPI-IPN, Mexico.

© *A. Gelbukh, C. Yáñez Márquez, O. Camacho Nieto (Eds.)*
Advances in Artificial Intelligence and Computer Science
Research on Computing Science 14, 2005, pp. 149-155

with a modified version of Mutual Information, in which the total amount of Web-pages delivered by the engine is replaced by the occurrence number of the most frequent word in language.

We know at least the following reasons for word connectivity evaluations. The first is the need of computational linguistics to extract from texts the stable word combinations — collocations and coordinate pairs. Being gathered into special DBs (e.g., [8], collocations can be used in diverse applications. The second reason is direct use of connectivity measurements for detection and correction of malapropisms, i.e. semantic errors of special type. It is shown that erroneous word combinations always have *SCI* values lesser than the intended (correct) combinations [2]. The third reason is selection of sequences of words — composite terms and names — that are to be used in information retrieval.

For experiments in Russian, Yandex search engine was used [11].

2. Word Connectivity

Each natural language text is a sequence of *word forms*, i.e. strings of letters from one delimiter to the next (e.g., *links*, *are*, *very*, *short*). Word forms pertaining to the morpho-paradigm with common meaning are associated into lexemes. One word form from a paradigm is taken as the lexeme's title for the corresponding dictionary entry, e.g. ***pen*** is taken for {*pen, pens*}; ***go***, for {*go, going, gone, went*}. In languages with rich morphology, for example, for verbs in Spanish, morpho-paradigms are broader.

We divide word forms into three categories:

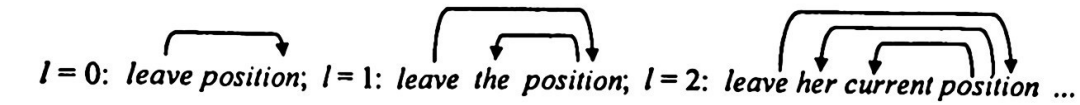
- ***Content words***: nouns; adjectives; adverbs; verbs except auxiliary and modal ones;
- ***Functional words***: prepositions; coordinate conjunctions; auxiliary and modal verbs;
- ***Stop words***: pronouns; proper names except of well known geographic or economic objects or personalities reflected in academic dictionaries and encyclopedias (these are considered content words); any other POS.

According to Dependency Grammars [7], each sentence can be represented at the syntactic level as a dependency tree with directed links "head → its dependent" between word-form labeled nodes. Following these links in the same direction of the arcs, from one content node through any linking functional nodes down to another content node, we get labeled subtree structure corresponding to a word combination. In a sensible text, we may consider each revealed combination with subordinate dependencies as a collocation. E.g., in the sentence *she hurriedly went through the big forest*, the collocations are ***went*** → *through* → ***forest***, ***hurriedly*** ← ***went*** and ***big*** ← ***forest***, whereas *she* ← ***went*** and *the* ← ***forest*** are not collocations, as having stop words at the extremes. The combinations may be also of coordinate type, e.g. ***mom*** → *and* → ***dad***. Thus, the syntactic links in word combinations can be immediate or realized through functional words.

We name the defined above type of word combinability lexico-syntactic connectivity. It implies a syntactic link between components and semantic compatibility of

corresponding lexemes. Such combinations are either idiomatic (the meaning of the whole is not equal to the sum of the component meanings) or free (the meaning of the whole is compositional). As to their stability, thus far not touched upon, it is very important feature for any applications: we should primarily work with stable word combinations in computational linguistics and information retrieval.

Combinable components can be linearly separated not only by their own functional word(s) but by many others dependent on one of the components. In other words, a close context in a dependency tree is in no way a close linear context. This makes difference with intensively studied bigrams. For example, collocation *to leave position* can contain intermediate contexts of any length *l*:



Coordinate pairs are word combination of two content words (or content word compounds) linked by a coordinative conjunction. In the most frequent case the components P_1 and P_2 of a stable coordinate pair are linked according to the formula $P_1 \rightarrow C \rightarrow P_2$, where the coordinate conjunction C is *and*, *or*, *but*.

The third type of word combinations interesting mainly for information retrieval is composite proper names. Many of them contain two word forms and can be treated as collocations (*President* → *Bush, George* ← *Bush*) or stable coordinate pairs (*Trinidad* → *and* → *Tobago*). However, many names contain three and more words. For humans, the composite names can contain addressing (*Sir*, *Mr.*, etc.), personal name(s), family name (usually repeating father's family name), patronymic name (derived from the father's personal name—in Russian tradition: *Boris Nikolayevich Yeltsyn*), family name of the mother (in Hispanic tradition: *Andrés López Obrador*). The binary decomposition necessary for connectivity evaluations is not clear for such sequences. However the usage of various shorter versions of names suggests corresponding dependency subtrees. For example, we take a subtree for the name of the VIP in the shape

Since one cannot say *President George*, the uniquely possible binary decomposition of the triple at the highest level is that shown by the vertical bar.

3. Numerical Criteria of Word Connectivity

Let us now consider occurrences of words and their co-occurrences as word combinations in a text corpus at some limited distance between them as random events. Their co-occurrence should be considered steady, or stable, if the relative frequency (=empirical probability) $N(P_1,P_2)/S$ of the co-occurrence of P_1 and P_2 is greater than the product of the relative frequencies $N(P_1)/S$ and $N(P_2)/S$ of the components taken apart (S is the corpus size). Using logarithms, we have the criterion of word connectivity well known as Mutual Information [10]:

$$MI(P_1,P_2) \equiv \log \frac{S \cdot N(P_1,P_2)}{N(P_1) \cdot N(P_2)}.$$

MI has important feature of scalability: if the sizes of all its 'building blocks' S, $N(P_1)$, $N(P_2)$, and $N(P_1,P_2)$ are multiplied by the same positive factor, *MI* conserves its value.

Other criteria different from *MI* but including the same building blocks is the scalable Pearson Correlation Coefficient [3]:

$$PCC(P_1,P_2) = \frac{\frac{S \cdot N(P_1,P_2)}{N(P_1) \cdot N(P_2)} - 1}{\sqrt{\left(\frac{S}{N(P_1)} - 1\right)\left(\frac{S}{N(P_2)} - 1\right)}}.$$

It scarcely can be more reasonably grounded than *MI*. As to Association Factor [10], it is not scalable, so its use for Web measurements is very doubtful. We ignore them both in this paper.

Any Web search engine automatically delivers statistics about a queried word or a word combination measured in numbers of Web-pages containing the event, and no information on word occurrences and co-occurrences is available. We then can re-conceptualize *MI* with all $N()$ as numbers of relevant pages and S as the page total managed by the engine. However, now $N()/S$ are not the empirical probabilities of relevant events: the same words entering a page are indistinguishable in the raw statistics, being counted only once, while the same page is counted repeatedly for each word included. We only keep a vague hope that the ratios $N()/S$ are monotonically connected with the corresponding empirical probabilities for words.

In such a situation we fill free to construe a new criterion from the same building blocks. Since evaluation of the page total S is multistage and not simple [1], we try to avoid its use in the target criterion but to conserve its scalability. The following criterion of word connectivity named by us Stable Connection Index seems good:

$$SCI(P_1,P_2) \equiv 16 + \log_2 \frac{N(P_1,P_2)}{\sqrt{N(P_1) \cdot N(P_2)}}.$$

The additive constant 16 and the logarithmic base 2 are chosen so that a great many of Russian word combinations intuitively considered cohesive fall into the interval 0 to 16. Hereafter we determine cohesive words P_1 and P_2 by the formula

$$SCI(P_1,P_2) > 0.$$

Depending on a specific search engine, the values $N(P_1)$, $N(P_2)$, and $N(P_1,P_2)$ can be got by only query (the case of Yandex for Russian language) or by three sequential queries close in time (the case of Google for English and Spanish). Anyhow, the scalability spares *SCI* of the influence of certain slow growth of the engine's resources. However, quick fluctuations of measurements from one access to Web to another implied by variations of trajectories within the engine's resources can be automatically compensated only in the case of obtaining all values through one query (Yandex). Fortunately, the quick fluctuations usually do not exceed ±5% of the measured values, and this gives the insignificant *SCI* variations (±0.1). By the way, it means that we should retain only one decimal digit after the point in *SCI* values.

Replacing S in *MI* by Web-page number N_{max} valid for one of the most frequent functional words in a given language, we have also construed the criterion, which is very simi-

lar to *MI* and keeps the scalability. We name it Modified Mutual Information and intend it for comparison with *SCI*:

$$MMI(P_1, P_2) \equiv k_1 \log_2 \frac{k_2 \cdot N_{\max} \cdot N(P_1, P_2)}{N(P_1) \cdot N(P_2)}$$

The constants k_1 and k_2, and the functional word met in $N_{\max}$ pages will be chosen later.

4. An Experimental Set and Comparison of Criteria

Our rather large experimental set for *SCI* evaluations is a collection of ca. 2200 Russian coordinate pairs. Similar pairs exist in any European language. This can be demonstrated by the following Russian stable coordinate pairs: *damy i gospoda* 'ladies and gentlemen'; *žaloby i predloženiya* 'complaints and suggestions'; *geodezija i kartografija* 'geodesy and cartography'; *avtobusy i avtomobili* 'buses and cars'; *amerikanskij i britanskij* 'American and British'; *analiz i prognoz* 'analysis and forecasting'; *bezopasnost' i obščestvennyj porjadok* 'security and social order'; *biznes i vlast'* 'business and authorities,' *etc.* Many of these pairs are sci-tech, economical or cultural terms and thus can be used for information retrieval.

SCI values were computed for all these pairs. In overwhelming majority they proved to be positive, so they are stable. The distribution of *SCI* values rounded to the nearest integers is given in Fig. 1. It has concentrated bell-like form with the mean value $M_{SCI} = 7.30$ and the standard deviation $D_{SCI} = 3.23$. As many as 69% *SCI* values are in the interval $M_{SCI} \pm D_{SCI}$.

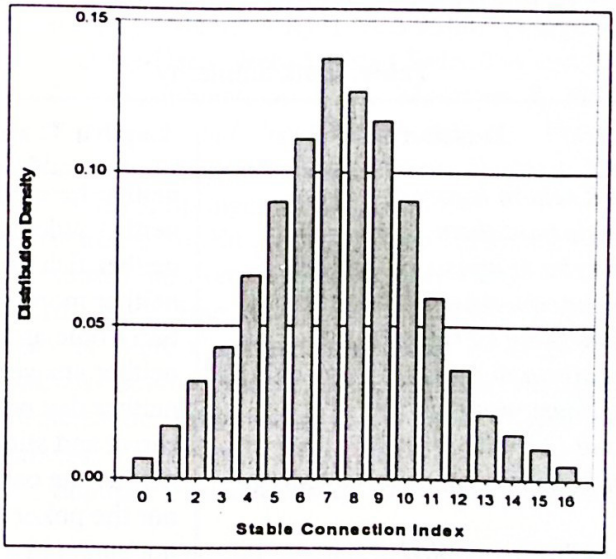


Fig. 1. Distribution for Stable Connection Index

Another criterion computed for the same set was Modified Mutual Information. We have selected Russian functional word *i* 'and' as having the rank 2 among the most frequent Russian words. During the experiment we observe $N_{\max} \approx 1.5 \cdot 10^9$, and constants k_1 and k_2 were chosen so that the mean values and the standard deviations for both criteria were nearly the same: $k_1 = 0.7$ and $k_2 = 360$ gave $M_{MMI} = 7.09$ and $D_{MMI} = 3.32$.

The direct comparison of *SCI* and *MMI* distributions (Fig. 1 and Fig. 2) shows their proximity. Making a difference, the *SCI* distribution has strict cutoff edge 16, while the *MMI* distribution is sloping more gently to the greater values. Nevertheless, the computing of the cosine value between the two vectors of measurements

$$\cos(SCI, MMI) = \frac{\sum_i (SCI_i \cdot MMI_i)}{\sqrt{\sum_i SCI_i^2 \sum_i MMI_i^2}}$$

gave the value .96, thus demonstrating nearly complete coincidence of the criteria.

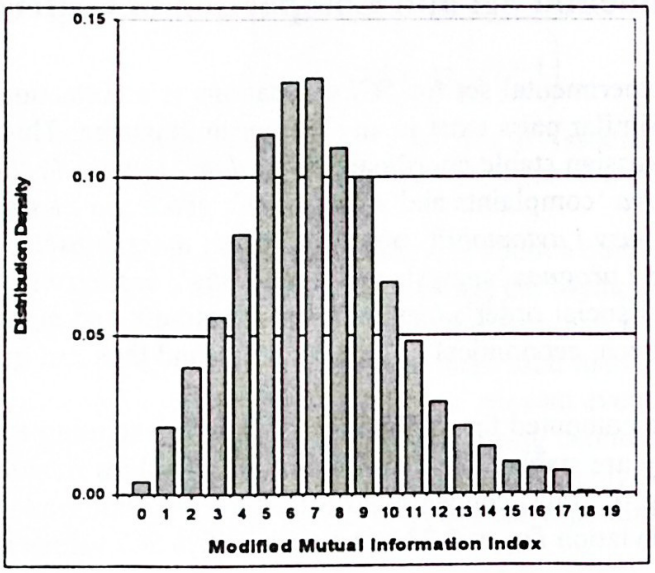


Fig. 2. Distribution for Modified Mutual Information

Table. Rank Similarity

SCI rank	*MMI* rank	Russian SCP	English Translation
1	12	*ni k selu ni k gorodu*	neither here nor there
2	21	*ni sluxu ni duxu*	neither hide nor hair
3	22	*ni ryba ni mjaso*	neither fish nor fowl
4	44	*ni mnogo ni malo*	neither more nor less
5	10	*San-Tome i Prinsipi*	San Tome and Principe
6	24	*ni otveta ni priveta*	neither answer nor regard
7	32	*ni dnem ni noč'ju*	neither day nor night
8	13	*knut i prjanik*	carrot and stick
9	2	*ni bogu svečka, ni čertu kočerga*	neither the candle for the God nor the poker for the devil
10	8	*to v žar, to v xolod*	hot and cold
11	28	*denno i noščno*	day and night
12	17	*meždu molotom i nakoval'nej*	between the beetle and the block
13	29	*vkriv' i vkos'*	Indiscriminately
14	52	*črezvyčajnyj i polnomočnyj*	Extraordinary and Plenipotentiary
15	27	*bez sučka i zadorinki*	without a hitch

Let us now compare the *SCI* and *MMI* values ranks for the same subset. Naturally that they proved to be different: some interspersion occurs. E.g., the initial fifteen *SCI* ranks differ from corresponding *MMI* ranks by 2 to 40. However, the maximal rank difference is

equal to only 1/50 of the whole set size. Hence we can conclude that the compared two criteria are nearly equivalent. The only reason to prefer *SCI* is that *MMI* requires four time-consuming accesses to the Web, while *SCI* needs only three of them.

5. Conclusions and Future Work

A convenient numerical measure for lexico-syntactic connection between content words is proposed — Stable Connection Index. It is computed based on the raw statistics automatically delivered by a Web search engine about pages with content words and their pairs. *SCI* is nearly insensitive to both slow growth of search engine's resources and quick fluctuations of raw Web statistics.

For comparison, another criterion of lexico-syntactic connection was also studied — Modified Mutual Information. It proved to be nearly equivalent to *SCI*, and the reason to prefer *SCI* is that *MMI* requires an additional access to the Web.

Both criteria can be used for acquisition of new collocations from the Web, for detection and correction of semantic errors in texts, and for extraction of composite terms and names needed in information retrieval.

References

1. Bolshakov, I.A., S.N. Galicia-Haro. Can We Correctly Estimate the Total Number of Pages in Google for a Specific Language? In: A. Gelbukh (Ed.). *Computational Linguistics and Intelligent Text Processing.* Proc. 4th Intern. Conf. on Computational Linguistics CICLing-2003, February 2003, Mexico City. LNCS 2588, Springer, 2003, p. 415-419.
2. Bolshakov, I.A., S.N. Galicia-Haro. Web-Assisted Detection and Correction of Joint and Disjoint Malapropos Word Combinations. In: A. Montoyo, R. Muñoz, E. Metais (Eds.) *Natural Language Processing and Information Systems.* Proc. 10th Intern. Conf. on Applications of Natural Language to Information Systems NLDB'2005, Alicante, Spain, June 2005. LNCS 3513, Springer, 2005, p.126-137.
3. Borko, H. The Construction of an Empirically Based Mathematically Derived Classification System. Proc. Western Joint Computer Conf., May 1962.
4. Gelbukh, A., G. Sidorov, S.-Y. Han, and E. Hernández-Rubio. Automatic Enrichment of Very Large Dictionary of Word Combinations on The Basis of Dependency Formalism. *Lecture Notes in Computer Science N 2972*, 2004, Springer-Verlag, pp 430-437.
5. Kilgarriff, A., G. Grefenstette. Introduction to the Special Issue on the Web as Corpus. *Computational linguistics*, V. 29, No. 3, 2003, p. 333-347.
6. Manning, Ch. D., H. Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, 1999.
7. Mel'čuk, I. *Dependency Syntax: Theory and Practice.* SUNY Press, NY, 1988.
8. *Oxford Collocations Dictionary for Students of English.* Oxford University Press, 2003.
9. Smadja, F. Retreiving Collocations from text: Xtract. *Computational Linguistics.* Vol. 19, No. 1, 1990, p. 143-177.
10. Stiles, H. E. The Association Factor in Information Retrieval. *Journal of the Association for Computing Machinery*, Vol. 8, No. 2, April 1961, pp. 271-279.
11. *Yandex*, http://www.yandex.ru