Distribution-Based Semantic Similarity of Nouns*

Igor A. Bolshakov and Alexander Gelbukh

Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan Dios Bátiz s/n, Col. Nueva Industrial Vallejo,
07738, Mexico City, Mexico
igor@cic.ipn.mx, gelbukh@gelbukh.com,
www.Gelbukh.com

Abstract. In our previous work we have proposed two methods for evaluating semantic similarity / dissimilarity of nouns based on their modifier sets registered in Oxford Collocation Dictionary for Student of English. In this paper we provide further details on the experimental support and discussion of these methods. Given two nouns, in the first method the similarity is measured by the relative size of the intersection of the sets of modifiers applicable to both of them. In the second method, the dissimilarity is measured by the difference between the mean values of cohesion between a noun and the two sets of modifiers: its own ones and those of the other noun in question. Here, the cohesion between words is measured via Web statistics for co-occurrences of words. The two proposed measures prove to be in approximately inverse dependency. Our experiments show that Web-based weighting (the second method) gives better results.

Keywords: Semantic relatedness, word space model, lexical resources, Web as corpus, natural language processing.

1 Introduction

Several works evaluate semantic similarity or dissimilarity between words, see [3, 11] and references therein. The majority of evaluations are based on semantic hierarchies of WordNet [4, 5]. In this class of methods, semantic dissimilarity between words is considered proportional to the number of steps separating corresponding nodes of the hierarchy. The nodes are synsets that include the words under evaluation, while the arcs are subset-to-superset links between the synsets. The greater the distance, the greater dissimilarity. This measure proved to be useful in many applications and tasks of computational linguistics, such as word sense disambiguation [9], information retrieval, etc.

Another possible way for estimation of semantic proximity of words consists in comparing the sets of other words frequently co-occurring in texts in close vicinity to the two words in question [6]. The more similar the recorded beforehand sets of standard neighbors of any two words of the same POS, the more semantically similar the

* Work done under partial support of Mexican Government (CONACyT, SNI, SIP-IPN, COTEPABE-IPN). Authors thank anonymous reviewers for valuable comments.

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 704–713, 2007.
© Springer-Verlag Berlin Heidelberg 2007

words [7]. As applied to nouns, the accompanying words are primordially their modifiers. In European languages, these are usually adjectives or participles; in English these are also nouns staying in preposition and used attributively.

We evaluate semantic similarity or dissimilarity of English nouns by two methods in this paper. Both of them are based on those standard modifier sets for few tens of commonly used English nouns that are registered for them in OCDSE that seems the most reliable source of English collocations so far [10]. The nouns were preferred with more numerous collections of modifiers recorded.

In the first method, the similarity $Sim(N_1, N_2)$ of the noun N_1 to the noun N_2 is measured by the ratio of the number of modifiers commonly applicable to the both nouns and the number of modifiers of N_2.

In the second method, we weight the relatedness between the noun and its modifiers by the Web co-occurrence statistics. Namely, the dissimilarity $DSim(N_1, N_2)$ of N_1 from N_2 is measured by the residual of two mean values of specially introduced *Stable Connection Index*. *SCI* is exteriorly like Mutual Information of two words [8] and operates by raw statistics of Web pages containing these words considered separately and in their close co-occurrences. In contrast to Mutual Information, it does not require repetitive evaluation of the total amount of pages under search engine's control. One mean value covers *SCI*s of all 'noun → its own modifier' pairs, another mean value covers *SCI*s of all 'N_1 → a modifier of N_2' pairs. English modifiers usually precede their nouns forming bigrams with them, thus facilitating reliable Web statistic evaluations. In other words, *Sim* is determined through coinciding modifiers of nouns, while *DSim* is determined through alien modifiers.

The main idea of the two methods discussed here was briefly presented in our previous work [2]. In this paper, we give more details on the experiments conducted to compare these two methods.

Namely, our experimental data show that though the *Sim* and *DSim* measures can be rather arbitrary in each specific case, on average they show an inverse monotonic interdependence. However, in our experiments *DSim* showed higher resolution. By higher resolution we mean that while many noun pairs have zero *Sim* values as measured according to the OCDSE, they differ significantly in their *DSim* values.

2 Modifier Sets Selected for Evaluations

English nouns with all their recorded modifiers—both adjectives and nouns in attributive use—were taken from OCDSE. The nouns were picked up in rather arbitrary manner, without taking into account their mental similarity. Our only preferences were with the nouns with larger modifier sets.

For 32 nouns taken, total amount of modifiers (partially repeating) is 1964, and the mean modifiers group size equals to 61.4, varying from 39 (for *comment* and *disease*) to 119 (for *eyes*). The second and the third ranks determined by the set sizes are with *expression* (115) and *effect* (105). The nouns selected and sizes of their modifier sets are shown in Table 1.

We have limited the number of nouns to 32 units, since the total amount of accesses to the Web in experiments of the second method (cf. Section 5) grows approximately as a square of the number of words in question, so that, taking into

Table 1. Selected nouns and sizes of their modifier sets

S/N	Noun	MSet Size	S/N	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

account limitations of Internet searchers and the general trend of all statistics to grow, we could afford several days to acquire all necessary statistics but not a month.

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

3 Influence of Intersection of Modifier Sets

In our first method, the similarity $Sim(N_i, N_j)$ is defined through the intersection ratio of modifier sets $M(N_i)$ and $M(N_j)$ of the two nouns by the formula

$$Sim\left(N_i, N_j\right) \equiv \frac{\left|M\left(N_i\right) \cap M\left(N_j\right)\right|}{\left|M\left(N_i\right)\right|}, \tag{1}$$

where $|M(N_i)|$ means cardinal number of the set $M(N_i)$, $\cap$ designates set intersection.

With such definition, the similarity measure is generally asymmetric: $Sim(N_i, N_j) \neq Sim(N_j, N_i)$, though both values are proportional to the number of commonly applicable modifiers. We can explain the asymmetry by means of the following extreme case. If $M(N_i) \subset M(N_j)$, each member of $M(N_i)$ has its own counterpart in $M(N_j)$, thus $Sim(N_i, N_j)$ reaches the maximum equal to 1 (just as when $M(N_i) = M(N_j)$), but some members of $M(N_j)$ have no counterparts in $M(N_i)$, so that $Sim(N_j, N_i) < 1$.

To better visualize the similarity, we put to Table 2 symmetric ratios

$$Sym\left(N_i, N_j\right) \equiv \frac{\left|M\left(N_i\right) \cap M\left(N_j\right)\right|}{\sqrt{\left|M\left(N_i\right)\right|\left|M\left(N_j\right)\right|}}.$$

Table 2. Similarity measure for the 32 nouns

2	9																													
3	5	7																												
4	2	7	3																											
5	10	2	4	2																										
6	7	2	9	2	5																									
7	6	11	7	2	11	6																								
8	10	2	2	4	11	4	4																							
9	4	6	0	7	6	2	13	4																						
10	12	5	11	0	0	7	9	2	6																					
11	5	9	12	10	6	9	12	1	3	8																				
12	6	10	28	2	2	18	4	4	2	11	14																			
13	0	0	2	2	3	0	2	2	0	0	2	4																		
14	0	4	6	0	8	4	8	5	3	0	5	14	2																	
15	0	2	2	6	7	6	0	2	0	0	9	2	0	4																
16	0	2	7	2	0	5	2	2	2	5	11	8	0	8	2															
17	9	3	20	3	3	3	4	1	0	3	7	13	2	1	0	8														
18	9	0	4	6	12	9	0	8	2	7	8	4	0	2	9	7	4													
19	9	6	6	4	0	9	4	2	6	10	9	12	0	2	0	6	12	11												
20	10	11	2	0	4	6	6	8	5	2	3	13	4	0	0	0	5	0	7											
21	10	2	7	2	5	12	6	6	6	11	8	5	2	3	5	3	2	7	4	9										
22	0	5	25	2	0	4	2	0	4	5	14	23	0	8	0	16	9	4	6	6	7									
23	4	4	3	2	4	2	4	11	4	4	6	4	4	2	2	0	5	0	8	8	6	2								
24	35	6	5	2	10	19	7	13	3	4	5	11	2	2	4	0	5	14	8	9	10	0	5							
25	8	1	3	3	7	6	5	5	4	1	5	5	3	2	1	0	5	6	10	9	5	0	6	12						
26	4	4	2	0	3	3	1	3	1	3	6	9	0	1	1	0	1	3	5	4	2	4	4	2	11					
27	3	2	1	3	2	2	3	8	4	2	3	1	2	0	0	0	1	2	4	4	1	0	3	4	12	9				
28	5	8	3	2	3	8	11	4	8	3	9	6	2	4	2	3	1	3	5	6	9	2	4	6	1	1	0			
29	8	2	5	2	2	10	4	7	3	2	8	4	0	3	2	2	4	2	6	7	7	2	0	10	1	1	5	4		
30	0	2	8	2	2	15	6	2	4	9	8	23	2	0	0	6	12	2	3	19	9	8	4	2	7	1	3	7	4	
31	0	7	5	2	0	0	0	0	2	2	1	2	0	0	0	0	1	0	0	0	0	2	0	0	1	3	1	2	0	2
32	4	6	5	2	2	2	4	4	5	2	7	6	2	4	0	4	4	2	5	8	3	0	4	7	8	4	7	4	2	10	0
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**

Because of the symmetry, we may consider the elements (in percents) only with $i \in [1, ..., j-1]$ (numbers at the left border and the bottom of the Table 2 are as in Table 1) and $j \in [1, ..., 32]$.

The average value of $Sym(N_i, N_j)$ in Table 1 is the 0.05. It reaches the maximum 0.35 for the pair {*answer*, *explanation*}, which are semantically closest words among these few under consideration. The following pairs decreasing in similarity are {*change, difference*}, {*change, expansion*}, {*difference, expansion*}, {*difference, feature*}, {*change, effect*}, {*comment, explanation*}, {*concept, difference*}, {*economy, expansion*}, {*concept, feature*}, etc. Some meaning clusters could be already seen, but we should admit that a more trustworthy clustering needs a vaster set of nouns and vaster sets of modifiers.

Quite dissimilar pairs (with zero *Sym* value) are quite numerous (76): {*change, copy*}, {*charge, decision*}, {*comment, decision*}, {*answer, disease*}, {*chance, disease*}, etc. The nouns for human features *eyes* and *face* proved to be very productive in modifiers (119 and 96 relatively) but very specific (their *Sym* measures are close to zero for majority of noun pairs).

4 Words Cohesion in Internet

Any words W_1 and W_2 may be considered forming a stable combination if their co-occurrence number $N(W_1, W_2)$ in a text corpus divided by S (the total number of words in the corpus) is greater than the product of relative frequencies $N(W_1)/S$ and $N(W_2)/S$ of the words considered apart. Using logarithms, we have a measure of word cohesion known as log-likelihood ratio or Mutual Information [8]:

$$MI(W_1, W_2) \equiv \log \frac{S \cdot N(W_1, W_2)}{N(W_1) \cdot N(W_2)}.$$

MI has important feature of scalability: if the values of all its building blocks S, $N(W_1)$, $N(W_2)$, and $N(W_1, W_2)$ are multiplied by the same factor, *MI* preserves its value.

Any Web search engine automatically delivers statistics on a queried word or a word combination measured in numbers of relevant Web pages, and no direct information on word occurrences or co-occurrences is available. We 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 corresponding events: the words that occur at the same a page are indistinguishable in the raw statistics, being counted only once, and the same page is counted repeatedly for each word included. We only hope that the ratios $N()/S$ are monotonically connected with the corresponding empirical probabilities for the events under consideration.

In such a situation a different word cohesion measure was construed from the same building blocks [1]. It conserves the feature of scalability, gives very close to *MI* results for statistical description of rather large sets of word combinations, but at the same time is simpler to be reached, since does not require repeated evaluation of the whole number of pages under the searcher's control. The new cohesion measure was named Stable Connection Index:

$$SCI(W_1, W_2) \equiv 16 + \log_2 \frac{N(W_1, W_2)}{\sqrt{N(W_1) \cdot N(W_2)}}. \quad (2)$$

The additive constant 16 and the logarithmic base 2 were chosen rather arbitrary. The constant 16 does not affect the comparisons discussed in this paper and is included purely for sake of tradition (since this is how the notion of *SCI* has been introduced previously); the reader can safely ignore it.

Since our experiments with Internet searchers need minimally several days to perform, some additional words on Web searchers are worthwhile here.

The statistics of searcher have two sources of variation in time. The first one is monotonic growing because of steady enlargement of searcher's DB. In our experience, for huge searcher's BDs and the queried words forming stable combinations, the

raw statistics $N(W_1)$, $N(W_2)$, $N(W_1,W_2)$ grow approximately with the same speed, so that *SCI* keeps its value with the precision to the second decimal digit, even if the statistics are got in different moments along the experimental day.

The second, fluctuating source of instability of Internet statistics is selection by the searcher of a specific processor and a specific path through searcher's DB—for each specific query. With respect to this, the searchers are rather different. For example, Google, after giving several very close statistics for a repeating query, can play a trick, suddenly giving twice fewer amount (with the same set of initial snippets!), thus shifting *SCI* significantly. Since we did not suffer of such troubles so far on behalf of AltaVista, we preferred it for our experiments.

5 Dissimilarity Based on Mean Cohesion Values

Consider first the mean cohesion values

$$\frac{1}{|M(N_i)|}\sum_{x\in M(N_i)} SCI(N_i,x)$$

between the noun N_i and all modifiers in its own modifier set $M(N_i)$. One can see in Table 3 that all mean *SCI* values are positive and mainly rather big (4 to 8), except for *enquiries.* As to the latter, we may suppose that occurrence statistics of British National Corpus—the base for selection of collocations in OCDSE—differ radically from Internet statistics, probably because OCDSE is oriented to the British variant of the English language, while Internet is mostly composed of texts written in American English or in international sort of English. Hence the collocations *intellectual / joint / open / critical / sociological... enquiries,* being rather rare in whole Internet, were inserted to OCDSE by purely British reasons. This is not unique case of British vs. USA language discrepancies. We had rejected orthographic differences like *flavour* vs. *flavor*, but we did not feel free to sift out such OCDSE collocations as *coastal flat* 'property by the sea,' which proved to be rare in Internet as a whole.

When calculating the SCI value of 'noun → modifier of a different noun' pairs that mainly are not normal collocations, we frequently observe the cases with zero co-occurrence number in Internet. Then formula (2) gives *SCI* value equal to $-\infty$. To avoid the singularity, we take the value –16 for such cases, i.e. the maximally possible positive value, but with the opposite sign.

We define the dissimilarity measure as

$$DSim(N_i,N_j)=\frac{1}{|M(N_i)|}\sum_{x\in M(N_i)}\left(SCI(N_i,x)-SCI(N_j,x)\right) \tag{3}$$

i.e., as the mean difference between the *SCI* value of the modifiers of N_i with N_i and N_j, respectively. Note that in this formula the noun in question is compared with the set of its own modifiers defined by the dictionary and with the set of the modifiers of the other noun. Two things can be observed as to this definition.

Table 3. The mean SCI values of nouns with their own modifiers

S/N	Noun	Mean SCI	S/N	Noun	Mean SCI
1	*answer*	6.3	17	*effect*	6.7
2	*chance*	4.9	18	*enquiries*	1.4
3	*change*	6.5	19	*evidence*	8.0
4	*charge*	5.6	20	*example*	6.1
5	*comment*	4.4	21	*exercises*	4.0
6	*concept*	5.9	22	*expansion*	6.4
7	*conditions*	6.5	23	*experience*	7.7
8	*conversation*	6.0	24	*explanation*	6.1
9	*copy*	5.4	25	*expression*	4.9
10	*decision*	7.2	26	*eyes*	6.0
11	*demands*	4.1	27	*face*	5.7
12	*difference*	6.2	28	*facility*	4.5
13	*disease*	8.3	29	*fashion*	5.1
14	*distribution*	6.7	30	*feature*	5.9
15	*duty*	5.6	31	*flat*	4.3
16	*economy*	6.7	32	*flavor*	6.1

First, the formula is not symmetric. As it was discussed above, we consider the relations between different nouns more as inclusion than as distance: *cat* is a perfect *animal*, i.e., in our terminology we would say that *cat* is no different from *animal*, while *animal* by no means is a perfect *cat*.

Another observation about this definition is more theoretical. It seems to be contradicting: while we use the objective reality, the Web (as corpus) to measure the relatedness between a noun and a modifier, we seemingly arbitrary restrict the set of participating modifiers to be considered by those found in a dictionary, which were subjectively selected by a lexicographer. What is more, this seemingly leads to the necessity to use in our method a specialized large lexical resource, which does not exist in all languages, and it is not clear how the results obtained with different such resources would coincide.

Though we did not conduct any corresponding experiments, we believe that the formula above can be modified to use the whole set of words of the language (occurring in a large corpus or in the Web). The formula is then to be modified to take into account the cohesion between each word and the noun in question; those words that have low value of such cohesion would be weighted out. However, this would be a bit impractical. So we here use an approximation to such a totally unsupervised approach. Our approximation takes advantage of an already existing resource to roughly indicate which words are expected to correlate with the given noun.

Note that in this sense the second method can be thought of as a weighted variant of the first one.

Table 4 shows the pairs with the smallest and the greatest dissimilarity measure in our small dataset. One can notice the pairs with the smallest dissimilarity, such as {*enquiries, explanation*}, do have similar or related meaning, while those with greater dissimilarity, such as {*disease, enquiries*}, look totally unrelated.

Table 4. Most and least similar noun pairs in our sample

Least dissimilar noun pairs				Most dissimilar noun pairs			
$Noun_1$	**$Noun_2$**	**DSim**	**Sim**	**$Noun_1$**	**$Noun_2$**	**DSim**	**Sim**
enquiries	*explanation*	0.3	0.156	*disease*	*enquiries*	18.5	0.000
enquiries	*distribution*	0.5	0.022	*eyes*	*enquiries*	15.8	0.017
enquiries	*comment*	0.6	0.111	*effect*	*enquiries*	14.8	0.029
enquiries	*conversation*	0.6	0.089	*face*	*enquiries*	14.7	0.010
enquiries	*change*	0.9	0.044	*experience*	*enquiries*	14.4	0.000
difference	*change*	1.1	0.321	*disease*	*economy*	14.2	0.000
enquiries	*fashion*	1.1	0.022	*disease*	*chance*	14.0	0.000
enquiries	*charge*	1.2	0.067	*flavor*	*enquiries*	14.0	0.020

In fact, the very small DSim measure can indicate that the words are nearly synonyms or nearly antonyms, but this results from a different our research.

6 Comparison and Discussion

Comparing the *Sim* and *DSim* values for the 16 pairs in Table 4, one can see that the pairs with maximal *Sim* values usually have minimal *DSim* values and vice versa, i.e. an inverse monotonic dependency exists between the two measures. More representative comparison is given in Figure 1 that gives correlations between *Sim* and *DSim* on the plane.

A statistically proved inverse monotonic dependency is quite clear form Figure 1. One can also comprehend that *DSim* has higher resolution for semantically most different nouns. Indeed, the numerous pairs with zero *Sim* values have quite diverse *DSim* values, from 14.0 for {*disease, flat*} to 4.2 for {*flat, answer*}. Hence the use of *DSim* measure seems preferable.

7 Conclusions and Future Work

Two methods of numerical evaluation of semantic similarity of any nouns is proposed. The evaluations are based on comparison of standard modifiers of the nouns registered in OCDSE. The first method evaluates similarity by the portion of common modifiers of the nouns, while the second one evaluates dissimilarity by the change of the mean cohesion of a given modifier set with its own noun and an alien one.

Cohesion measurements are based on raw Web statistics of occurrences and co-occurrences of supposedly cohesive words. It is shown that dissimilarity measured through the Web has higher resolution and thus may have greater reliability.

Both methods do not depend on language and can be easily tested on the resources of other languages. Currently we are conducting experiments with Spanish and Russian, which are morphologically-rich languages. For English, it is worthwhile to repeat evaluations for a greater number of nouns and for different source of modifiers sets, e.g. for a large corpus of American origin. Finally, we believe that this method can be applied to words of parts of speech other than nouns, though one should be much more careful with, say, verbs, where the co-occurrence patterns are much more lexicalized and less semantic than those of nouns.

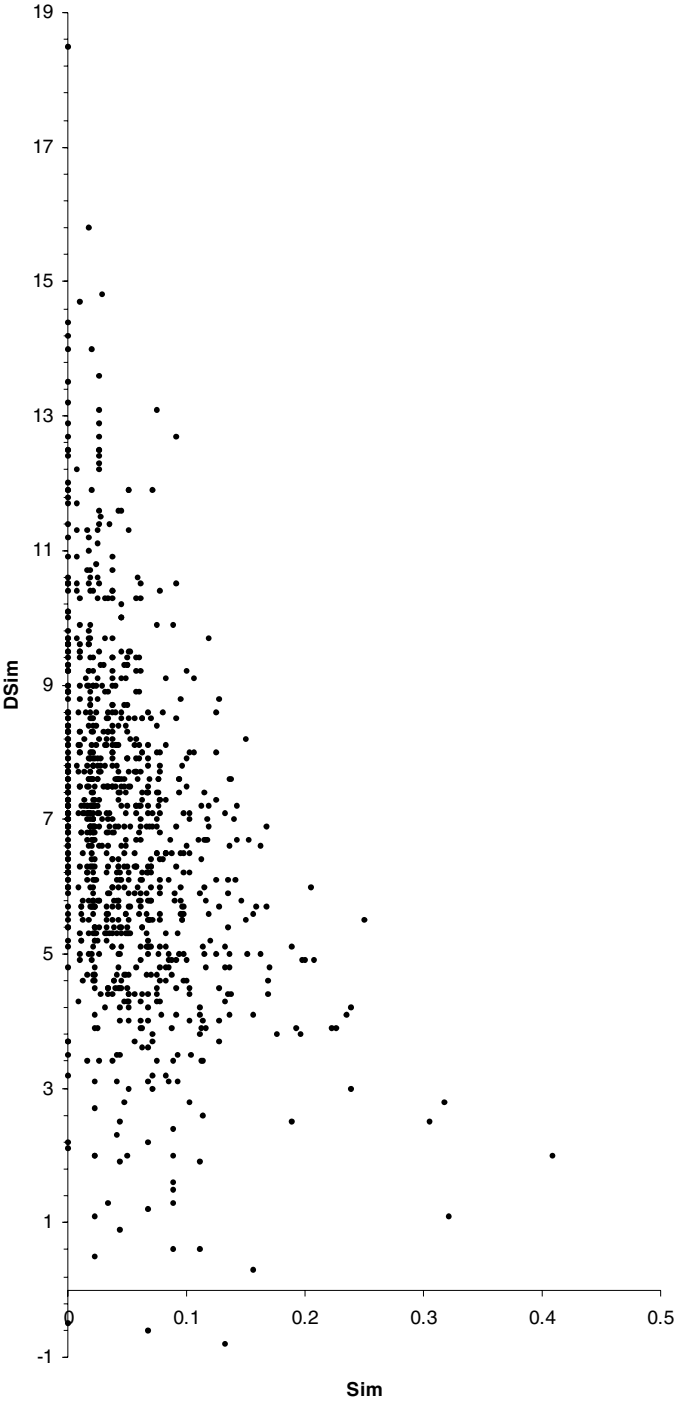

Fig. 1. Correlations between *Sim* and *DSim*

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. Bolshakov, I.A., Gelbukh, A.: Two Methods of Evaluation of Semantic Similarity of Nouns Based on Their Modifier Sets. In: LNCS, vol. 4592, Springer, Heidelberg (2007)
3. Cilibrasi, R.L., Vitányi, P.M.B.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007), www.cwi.nl/ paulv/ papers/tkde06.pdf
4. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
5. Hirst, G., Budanitsky, A.: Correcting Real-Word Spelling Errors by Restoring Lexical Cohesion. Natural Language Engineering 11(1), 87–111 (2005)
6. Keller, F., Lapata, M.: Using the Web to Obtain Frequencies for Unseen Bigram. Computational linguistics 29(3), 459–484 (2003)
7. Lin, D.: Automatic retrieval and clustering of similar words. In: COLING-ACL 1998, Canada (1998)
8. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
9. McCarthy, D., Koeling, R., Weeds, J., Carroll, J.: Finding Predominant Word Senses in Untagged Text. In: Proc. 42nd Annual Meeting of the ACL, Barcelona, Spain (2004)
10. Oxford Collocations Dictionary for Students of English. Oxford University Press (2003)
11. Patwardhan, S., Banerjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, Springer, Heidelberg (2003)