Two Methods of Evaluation of Semantic Similarity of Nouns Based on Their Modifier Sets*

Igor A. Bolshakov, Alexander Gelbukh

Center for Computing Research (CIC),
National Polytechnic Institute (IPN), Mexico City, Mexico
{igor,gelbukh}@cic.ipn.mx

Abstract. Two methods of evaluation of semantic similarity/dissimilarity of English nouns are proposed based on their modifier sets taken from Oxford Collocation Dictionary for Student of English. The first method measures similarity by the portion of modifiers commonly applicable to both nouns under evaluation. The second method measures dissimilarity by the change of the mean value of cohesion between a noun and modifiers, its own or those of the contrasted noun. Cohesion between words is measured by Stable Connection Index (*SCI*) based of raw Web statistics for occurrences and co-occurrences of words. It is shown that the two proposed measures are approximately in inverse monotonic dependency, while the Web evaluations confer a higher resolution.

1. Introduction

There are numerous works on evaluation of semantic similarity/dissimilarity between words, see [10] and references therein for a review. The majority of evaluations are based on semantic hierarchies of WordNet or EuroWordNet [2, 3]. Semantic dissimilarity between words is measured by the number of steps that separate corresponding nodes of the hierarchy. The hierarchy nodes are synsets including the words under evaluation, while the arcs are subset-superset links connecting these synsets. The greater is the distance, the lower is similarity. This measure proved to be useful in many applications and tasks of computational linguistics, such as word sense disambiguation [8, 6], information retrieval, etc.

In fact, there exists an alternative way to evaluate semantic similarity, namely through comparison of the sets of words frequently co-occurring in texts in close vicinity to words under evaluation. The more similar are the recorded beforehand sets of standard neighbors of any two words of the same POS, the more semantically similar are the words. As applied to nouns, the accompanying words are primordially modifiers, whose role in European languages is usually played by adjectives and—in English—also by attributively used nouns staying in preposition.

In this paper, semantic similarity/dissimilarity of English nouns is evaluated by two different methods, both based on those standard modifier sets for few tens of commonly used English nouns that are registered for them in OCDSE—the most reliable

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

source of English collocations [9]. The nouns were selected with preference to those with greater numbers 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, 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 close in its definition to Mutual Information of two words [7]. It operates by raw statistics of Web pages that contain these words and their close co-occurrences and does not require repetitive evaluation of the total amount of pages under search engine's control [4]. One mean value covers *SCI*s of all 'noun → its own modifier' pairs, another mean value covers *SCI*s of all 'N_1 → modifier of N_2' pairs. English modifiers usually stay just before their nouns forming bigrams with them, and this facilitates rather reliable Web statistic evaluations.

To put it otherwise, *Sim* is determined through coinciding modifiers of nouns, while *DSim* is determined through alien modifiers. As the main result, the *Sim* and *DSim* measures proved to be approximately connected by inverse monotonic dependency. However, *DSim* seems preferable because of its higher resolution: the numerous noun pairs with zero *Sim* values differ significantly with respect to *DSim*.

2. Experimental Modifier Sets

We took English nouns with all their recorded modifiers—both adjectives and nouns in attributive use—from OCDSE. The nouns were picked up from there in rather arbitrary manner, approximately one noun per nine OCDSE pages. At the same time, our preferences were with the most productive nouns, i.e. having vaster modifier sets.

Table 1. Selected nouns and sizes of their modifier sets

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

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 demonstrated in Table 1.

We had to limit 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 rapidly, approximately as (*Nouns* + 40) × (*Nouns* + 1), so that, taking into account severe limitations of Internet searchers, we could afford several days for acquiring all necessary statistics, but scarcely a month or more.

The nouns *conditions, demands, enquiries, exercises,* and *eyes* were taken in plural, since they proved to be more frequently used with their recorded modifier sets in plural than in singular.

3. Semantic Similarity Based on Intersection of Modifier Sets

The similarity $Sim(N_i, N_j)$ in the first method is mathematically 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 $M(N_i)$ and $\cap$ set intersection, cf. [6].

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 $Sim(N_j, N_i) < 1$.

4. Measurement of Words Cohesion by means of Internet

It is well-known that any two 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 obtain the log-likelihood ratio or Mutual Information [7]:

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

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 relevant events: the words that occur at the same 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 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 got from Internet, since does not require repeated evaluation of the whole number of pages under 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, but such scaling factors do not hamper the purposes of this paper and permit to consider words W_1 and W_2 cohesive, if $SCI(W_1, W_2)$ is positive.

Since our experiments with Internet searchers need at least several days to complete, some additional words on Web searchers are worthwhile now.

The statistics of searcher have two sources of changing in time. The first source is monotonic growing because of steady enlargement of searcher's DB. In our experience, for well saturated searcher's BDs and 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 the same value (with the precision to the second decimal digit), even if the statistics are got in different time along the day of experiments.

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 quite 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 purposes.

5. Semantic Dissimilarity Based on Mean Cohesion Values

Let us first consider the mean cohesion values

$$\frac{1}{|M(N_i)|} \sum_{A_k \subset M(N_i)} SCI(N_i, A_k)$$

between the noun N_i and all modifiers A_k in its own modifier set $M(N_i)$. One can see in the Table 2 that all mean SCI values are positive and mainly rather big (4 to 8), except for *enquiries*. On the latter occasion, we may suppose that occurrence statistics of British National Corpus—the base for selection of collocations in OCDSE—differ radically from Internet statistics that is not British oriented in its bulk. 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. Except of 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.

Table 2. The mean SCI values of nouns with their own modifiers

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

Passing to SCI evaluation of 'noun → modifier of a different noun' pairs that mainly are not normal collocations, we can frequently meet 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. maximally possible positive value, but with the opposite sign.

Table 3. Most and lest similar noun pairs

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

We determine the dissimilarity measure by the formula

$$DSim(N_i, N_j) = \frac{1}{|M(N_i)|} \sum_{A_k \in M(N_i)} SCI(N_i, A_k) - \frac{1}{|M(N_j)|} \sum_{A_k \in M(N_i)} SCI(N_j, A_k) \qquad (3)$$

The diminuend at the right part of (3) is the mean *SCI* value of N_i with its own modifiers, while the subtrahend is the mean *SCI* value of N_j estimated with respect to all modifiers of N_i. It is clear that $DSim(N_i, N_j)$ is minimal possible (0) for $i = j$.

For different nouns, lest and most dissimilar noun pairs are shown in Table 3. The pair {*enquiries, explanation*} proved to be the most similar by *DSim* criterion, while the pair {*disease, enquiries*}, the most dissimilar.

6. Conclusions

We have proposed two methods of how numerically evaluate semantic similarity of any two English nouns. The evaluations are based on comparison of standard modifiers of the nouns. 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 between a given noun and modifiers, when the set of its own modifiers commuted into the set of alien ones.

The comparison of *Sim* and *DSim* values for as few as 16 pairs in Table 3 shows that the pairs with maximal *Sim* usually have minimal *DSim* and vice versa, i.e. an inverse monotonic dependency exists between the two measures. One can note 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.

Cohesion measurements are based on raw Web statistics of occurrences and co-occurrences of supposedly cohesive words. For both methods, the standard modifier sets are taken from Oxford Collocations Dictionary for Students of English. It is shown that dissimilarity measured through the Web has higher resolution and thus may have greater reliability.

Both method do not depend on language and can be easily tested on the resources of other 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.

References

1. Bolshakov, I.A., E.I. Bolshakova. Measurements of Lexico-Syntactic Cohesion by means of Internet. *Lecture Notes in Artificial Intelligence*, N 3789, Springer, 2005, p. 790–799.
2. Fellbaum, Ch. (Ed.) *WordNet*: *An Electronic Lexical Database*. MIT Press, 1998.
3. Hirst, G., A. Budanitsky. Correcting Real-Word Spelling Errors by Restoring Lexical Cohesion. Natural Language Engineering, 11(1), 2005, 87–111.
4. Keller, F., M. Lapata. Using the Web to Obtain Frequencies for Unseen Bigram. *Computational linguistics*, V. 29, No. 3, 2003, p. 459-484.
5. Ledo-Mezquita, Y., G. Sidorov. *Combinación de los métodos de Lesk original y simplificado para desambiguación de sentidos de palabras*. International Workshop on Natural Language Understanding and Intelligent Access to Textual Information, in conjunction with MICAI-2005, Mexico, 2005, pp. 41-47.
6. Lin, Dekang. Automatic retrieval and clustering of similar words. *COLING-ACL 98*, 1998.
7. Manning, Ch. D., H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
8. McCarthy, Diana, Rob Koeling, Julie Weeds, John Carroll. Finding PredominantWord Senses in Untagged Text. ACL-2004.
9. *Oxford Collocations Dictionary for Students of English*. Oxford University Press, 2003.
10. Patwardhan, S., S. Banerjee, T. Pedersen. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Computational Linguistics and Intelligent Text Processing, CICLing-2003. *Lecture Notes in Computer Science*, N 2588, Spriner, 2003.