# Which statistics reflect semantics? Rethinking synonymy and word similarity

Derrick Higgins
Educational Testing Service
dhiggins@ets.org

## 1 Overview

A great deal of recent work addresses the task of statistical modeling of word similarity relations (cf.Schütze (1992), Lund and Burgess (1996) Landauer and Dumais (1997), Lin (1998), Turney (2001)). While this has largely been viewed as an engineering task (with the notable exception of much writing on Latent Semantic Analysis (LSA)), the relative success of different approaches to constructing word similarity measures is highly relevant to issues in theoretical semantics and language acquisition.

With this background in mind, this paper has two main aims. First, we present yet another statistical approach to the calculation of word-similarity scores (LC-IR), which significantly outperforms other methods on standard benchmarks including the 80-question set of TOEFL synonym test items first employed by Landauer and Dumais (1997). Second, we hope to demonstrate that

- various methods for assessing word similarity are based on fundamentally different assumptions about the statistical properties which synonyms can be expected to display,
- the performance of each method can be taken as a judgment on the validity of these assumptions, and
- whether these predictions regarding the statistical distribution of synonyms in a corpus are borne out ought to be taken into account in any consideration of the acquisition of meaning as part of language, and the mental representation of meaning.

# 2 LC-IR and its predecessors

Without indulging in too much of a caricature, we can classify different approaches to statistical estimation of word similarity according to the assumptions which they make about the distribution of synonyms (actually, *plesionyms*; cf. Edmonds and Hirst (2002)).

The techniques of Latent Semantic Analysis, Random Indexing, and HAL all collect statistics on the relative frequency with which a word appears "near" other words. Similar words can then be identified as those which have a similar profile of content words which tend to occur near them. The specifics vary between these different approaches to similarity calculation—for example, the proximity required for words to count as "near" one another varies from a distance of 3 words (Random Indexing) to as much as 300 words (LSA). Yet these approaches are similar enough that we can say they fundamentally depend on the assumption that *similar words tend to have the same neighboring content words*. We will refer to this as the **topicality** assumption, making the inference that synonyms tend to have the same neighbors because they are in passages which are on the same topic.

On the other hand, PMI-IR also involves the collection of statistics regarding the relative frequency with which word occur in proximity, but the assumption made regarding how this relates to synonymy is quite different. Instead of the assumption that similar words will occur near the same words, the calculation which forms the core of PMI-IR assumes that similar words will tend to occur near *each other*. The intuitive basis for this is not as clear as in the case of the topicality assumption, but the good results of PMI-IR lend it some empirical credence. We will refer to it as the **proximity** assumption.

Finally, Dekang Lin's 1998 work could be said to be based on the **parallelism** assumption: synonyms ought to be found in similar grammatical frames. The primary statistics gathered by Lin's method are the frequencies with which words occur linked by specific grammatical relations with other words.

Adding to this list of approaches, we present LC-IR (local context–information retrieval), a method for constructing word similarity scores which is inspired by PMI-IR, but which differs in its basic assumptions, and produces significantly better results. LC-IR, like PMI-IR, collects counts from the Web on how often words occur near one another, but it uses a smaller window size (requiring absolute adjacency). At first glance, this would seem to be a minor modification to the basic PMI-IR model, and not one which influences its fundamental assumptions. However, the small window size is of paramount importance to the model, and almost guarantees that LC-IR will identify synonyms conforming to the **parallelism** assumption, whereas standard PMI-IR is based on the more nebulous **proximity** assumption.

# 3 Applying LC-IR

As stated above, LC-IR is quite similar to PMI-IR: it also uses the AltaVista search engine as its basic data structure for establishing the similarity of words, and it also basically relies on statistics about how often words occur near one another. Turney's 2001 index of how similar two words are is given by the pointwise mutual information of the two words (1), which in the case of a web search using the NEAR operator can be estimated as in (2).[1] (This restatement amounts to giving a specific interpretation to the "&" operator in (1).)

$$\text{Similarity}_{\text{PMI-IR}}(w_1, w_2) = \frac{\text{p}(w_1 \ \& \ w_2)}{\text{p}(w_1)\text{p}(w_2)} \quad (1)$$

$$\approx \frac{\text{hits}(w_1 \text{ NEAR } w_2)}{\text{hits}(w_1)\text{hits}(w_2)} \quad (2)$$

On the other hand, LC-IR is based on the similarity metric in (3).[2] This metric differs from Turney's in two main ways. First, it uses the frequency with which words are found adjacent to one another, rather than the frequency with which they are found within a ten-word window of one another. This requirement, along with the fact that AltaVista's search engine ignores punctuation such as commas, curiously returns almost exclusively documents in which the two words in question are conjoined as part of a list. For example, a search for the exact phrase "rambunctious playful" on AltaVista returned 37 pages, most of which contained the phrase as part of a list of adjectives used to describe the same entity. By contrast, a search for the phrase "rambunctious artful" returned no documents. In essence, it turns out that LC-IR presumes a very strict version of the parallelism constraint discussed above. More than simply requiring that two words be used in the same grammatical frame, LC-IR asks that they actually be conjoined in the same sentence, resulting in syntactic and semantic parallelism.

$$\text{Similarity}_{\text{LC-IR}}(w_1, w_2) = \frac{\min(\text{hits}(w_1 \ w_2), \text{hits}(w_2 \ w_1))}{\text{hits}(w_1)\text{hits}(w_2)} \quad (3)$$

The second refinement to the PMI-IR model which is evident in Equation 3 is the fact that we take the minimum number of hits for the two possible orders in which the words could be found in a document. This is necessary because of the possibility that one word order could be a collocation, and thus have a higher frequency of occurrence

[1] This presentation ignores Turney's discounting of contexts in which the word *not* appears. While this does improve his results slightly, it is not central to his method and need not clutter our discussion here.

[2] In fact, we use a simple method of discounting, subtracting one from the number of web hits found on each search, but again, this is a subtlety.

Table 1: Comparison of word similarity results across three synonym tests

| | TOEFL | RDWP | ESL | Overall |
|---|---|---|---|---|
| Baseline | $\frac{20}{80}$ = 25% | $\frac{75}{300}$ = 25% | $\frac{12.5}{50}$ = 25% | $\frac{107.5}{430}$ = 25% |
| LSA | $\frac{51.5}{80}$ = 64.4% | – | – | – |
| Random Indexing | $\frac{54}{80}$ = 67.5% | $\frac{109.2}{300}$ = 36.4% | $\frac{19.6}{50}$ = 39.2% | $\frac{182.8}{430}$ = 42.5% |
| PMI-IR | $\frac{64.25}{80}$ = 80.0% | $\frac{216.83}{300}$ = 72.3% | $\frac{33}{50}$ = 66.0% | $\frac{314.08}{430}$ = 73.0% |
| LC-IR | $\frac{65}{80}$ = 81.3% | $\frac{224.33}{300}$ = 74.8% | $\frac{39}{50}$ = 78.0% | $\frac{329.33}{430}$ = 76.6% |
| Roget's Thesaurus | $\frac{63}{80}$ = 78.8% | $\frac{223}{300}$ = 74.3% | $\frac{41}{50}$ = 82.0% | $\frac{327}{430}$ = 76.0% |

than is to be expected given the words' meaning alone, and also because part-of-speech ambiguities can result in one of the two sequences having an elevated frequency because it coincides with a common noun-verb or adjective-noun combination.

The performance of LC-IR in identifying synonyms, as measured by the standard benchmarks of the TOEFL, ESL, and Reader's Digest synonym test sets, is the highest yet recorded, exceeding even the results of systems using lexical resources such as *Roget's Thesaurus* (Jarmasz and Szpakowicz, 2003), as shown in Table 1. This improvement in the model's results suggests that the strong parallelism assumption used by LC-IR is, in fact a strong predictor of synonymy, and that this metric is usable despite the sparsity of the data available on exactly how often word pairs are used in this narrowly parallel fashion.

The baseline model of Table 1 simply guesses randomly at each item, resulting in an expected accuracy of 25%, since each synonym item has four possible answers. The corpus-based approaches of LSA and Random Indexing, based on the topicality assumption of synonym distribution, fare much better than the baseline, but not nearly as well as PMI-IR. Jarmasz and Szpakowicz (2003)'s approach of using the distance between words in a thesaurus fares better still, and represented the best-performing model across test sets until the development of LC-IR, which now takes a narrow lead.

# 4 Implications for a theory of lexical semantics and acquisition

In addition to support from the empirical results of the previous section, we also wish to claim that methods based on the parallelism assumption make the greatest contribution to a realistic model of semantic acquisition.

First, we consider the problem of "one-shot" word learning (Yip and Sussman, 1998), and argue that this phenomenon is more easily modeled as a special case of learning from parallel word usage than as any corresponding process involving topicality or simple proximity. The crucial point is that the primary datum for parallelism-based word similarity is a linguistic construction whose attestation is often definitive, whereas the requisite data for the other approaches is typically much too sparse for one-shot learning.

Second, we consider the more general issue of children's acquisition of vocabulary, addressed by Landauer and Dumais (1997). In this domain as well, we argue that parallelism is a better cue to word similarity than either topicality or proximity, in part due to the nature of the primary linguistic data with which the child is confronted. In addition to data sparsity issues, which are in play here as well as in adult word learning, such approaches would find it difficult to deal with language data which consists of relatively short utterances, and can be lacking in topical coherence.

# References


Edmonds, P. and G. Hirst (2002). Near-synonymy and lexical choice. *Computational Linguistics*, **28**(2):105–144.

Jarmasz, M. and S. Szpakowicz (2003). Roget's thesaurus and semantic similarity. University of Ottawa ms.

Landauer, T. K. and S. T. Dumais (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**:211–240.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. 15th International Conference on Machine Learning*, pp. 296–304. Morgan Kaufmann, San Francisco, CA.

Lund, K. and C. Burgess (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavioral Research Methods, Instruments and Computers*, **28**(2):203–208.

Schütze, H. (1992). Dimensions of meaning. In *Proceedings of Supercomputing '92, Minneapolis.*, pp. 787–796.

Turney, P. D. (2001). Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. In *Proc. 12th European Conference on Machine Learning*, pp. 491–450.

Yip, K. and G. J. Sussman (1998). Sparse representations for fast, one-shot learning. MIT AI Lab Memo 1633, May 1988.