# Automatic Retrieval and Clustering of Similar Words

**Dekang Lin**
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada R3T 2N2
lindek@cs.umanitoba.ca

## Abstract

Bootstrapping semantics from text is one of the greatest challenges in natural language learning. We first define a word similarity measure based on the distributional pattern of words. The similarity measure allows us to construct a thesaurus using a parsed corpus. We then present a new evaluation methodology for the automatically constructed thesaurus. The evaluation results show that the thesaurus is significantly closer to WordNet than Roget Thesaurus is.

## 1 Introduction

The meaning of an unknown word can often be inferred from its context. Consider the following (slightly modified) example in (Nida, 1975, p.167):

(1) A bottle of *tezgüino* is on the table.
Everyone likes *tezgüino*.
*Tezgüino* makes you drunk.
We make *tezgüino* out of corn.

The contexts in which the word *tezgüino* is used suggest that *tezgüino* may be a kind of alcoholic beverage made from corn mash.

Bootstrapping semantics from text is one of the greatest challenges in natural language learning. It has been argued that similarity plays an important role in word acquisition (Gentner, 1982). Identifying similar words is an initial step in learning the definition of a word. This paper presents a method for making this first step. For example, given a corpus that includes the sentences in (1), our goal is to be able to infer that *tezgüino* is similar to "beer", "wine", "vodka", etc.

In addition to the long-term goal of bootstrapping semantics from text, automatic identification of similar words has many immediate applications. The most obvious one is thesaurus construction. An automatically created thesaurus offers many advantages over manually constructed thesauri. Firstly, the terms can be corpus- or genre-specific. Manually constructed general-purpose dictionaries and thesauri include many usages that are very infrequent in a particular corpus or genre of documents. For example, one of the 8 senses of "company" in WordNet 1.5 is a "visitor/visitant", which is a hyponym of "person". This usage of the word is practically never used in newspaper articles. However, its existance may prevent a co-reference recognizer to rule out the possiblity for personal pronouns to refer to "company". Secondly, certain word usages may be particular to a period of time, which are unlikely to be captured by manually compiled lexicons. For example, among 274 occurrences of the word "westerner" in a 45 million word San Jose Mercury corpus, 55% of them refer to hostages. If one needs to search hostage-related articles, "westerner" may well be a good search term.

Another application of automatically extracted similar words is to help solve the problem of data sparseness in statistical natural language processing (Dagan et al., 1994; Essen and Steinbiss, 1992). When the frequency of a word does not warrant reliable maximum likelihood estimation, its probability can be computed as a weighted sum of the probabilities of words that are similar to it. It was shown in (Dagan et al., 1997) that a similarity-based smoothing method achieved much better results than back-off smoothing methods in word sense disambiguation.

The remainder of the paper is organized as follows. The next section is concerned with similarities between words based on their distributional patterns. The similarity measure can then be used to create a thesaurus. In Section 3, we evaluate the constructed thesauri by computing the similarity between their entries and entries in manually created thesauri. Section 4 briefly discuss future work in clustering similar words. Finally, Section 5 reviews related work and summarize our contributions.

## 2 Word Similarity

Our similarity measure is based on a proposal in (Lin, 1997), where the similarity between two objects is defined to be the amount of information contained in the commonality between the objects divided by the amount of information in the descriptions of the objects.

We use a broad-coverage parser (Lin, 1993; Lin, 1994) to extract dependency triples from the text corpus. A dependency triple consists of two words and the grammatical relationship between them in the input sentence. For example, the triples extracted from the sentence "I have a brown dog" are:

(2) (have subj I), (I subj-of have), (dog obj-of have), (dog adj-mod brown), (brown adj-mod-of dog), (dog det a), (a det-of dog)

We use the notation $\|w, r, w'\|$ to denote the frequency count of the dependency triple $(w, r, w')$ in the parsed corpus. When $w$, $r$, or $w'$ is the wild card ($*$), the frequency counts of all the dependency triples that matches the rest of the pattern are summed up. For example, $\|cook, obj, *\|$ is the total occurrences of cook–object relationships in the parsed corpus, and $\|*, *, *\|$ is the total number of dependency triples extracted from the parsed corpus.

The description of a word $w$ consists of the frequency counts of all the dependency triples that matches the pattern $(w, *, *)$. The commonality between two words consists of the dependency triples that appear in the descriptions of both words. For example, (3) is the the description of the word "cell".

(3) ||cell, subj-of, absorb||=1
||cell, subj-of, adapt||=1
||cell, subj-of, behave||=1
.....
||cell, pobj-of, in||=159
||cell, pobj-of, inside||=16
||cell, pobj-of, into||=30
.....
||cell, nmod-of, abnormality||=3
||cell, nmod-of, anemia||=8
||cell, nmod-of, architecture||=1
.....
||cell, obj-of, attack||=6
||cell, obj-of, bludgeon||=1
||cell, obj-of, call||=11
||cell, obj-of, come from||=3
||cell, obj-of, contain||=4
||cell, obj-of, decorate||=2
.....
||cell, nmod, bacteria||=3
||cell, nmod, blood vessel||=1
||cell, nmod, body||=2
||cell, nmod, bone marrow||=2
||cell, nmod, burial||=1
||cell, nmod, chameleon||=1
.....

Assuming that the frequency counts of the dependency triples are independent of each other, the information contained in the description of a word is the sum of the information contained in each individual frequency count.

To measure the information contained in the statement $\|w, r, w'\|$=c, we first measure the amount of information in the statement that a randomly selected dependency triple is $(w, r, w')$ when we do not know the value of $\|w, r, w'\|$. We then measure the amount of information in the same statement when we do know the value of $\|w, r, w'\|$. The difference between these two amounts is taken to be the information contained in $\|w, r, w'\|$=c.

An occurrence of a dependency triple $(w, r, w')$ can be regarded as the co-occurrence of three events:

$A$: a randomly selected word is $w$;

$B$: a randomly selected dependency type is $r$;

$C$: a randomly selected word is $w'$.

When the value of $\|w, r, w'\|$ is unknown, we assume that $A$ and $C$ are conditionally independent given $B$. The probability of $A$, $B$ and $C$ co-occurring is estimated by

$$P_{\text{MLE}}(B)P_{\text{MLE}}(A|B)P_{\text{MLE}}(C|B),$$

where $P_{\text{MLE}}$ is the maximum likelihood estimation of a probability distribution and

$$\begin{aligned} P_{\text{MLE}}(B) &= \frac{\|*,r,*\|}{\|*,*,*\|}, \\ P_{\text{MLE}}(A|B) &= \frac{\|w,r,*\|}{\|*,r,*\|}, \\ P_{\text{MLE}}(C|B) &= \frac{\|*,r,w'\|}{\|*,r,*\|} \end{aligned}$$

When the value of $\|w, r, w'\|$ is known, we can obtain $P_{\text{MLE}}(A, B, C)$ directly:

$$P_{\text{MLE}}(A, B, C) = \|w, r, w\|/\|*, *, *\|$$

Let $I(w, r, w')$ denote the amount information contained in $\|w, r, w'\|$=c. Its value can be com-

$$\text{sim}_{Hindle}(w_1, w_2) = \sum_{(r,w) \in T(w_1) \cap T(w_2) \wedge r \in \{\text{subj-of, obj-of}\}} \min(I(w_1, r, w), I(w_2, r, w))$$

$$\text{sim}_{Hindle_r}(w_1, w_2) = \sum_{(r,w) \in T(w_1) \cap T(w_2)} \min(I(w_1, r, w), I(w_2, r, w))$$

$$\text{sim}_{cosine}(w_1, w_2) = \frac{|T(w_1) \cap T(w_2)|}{\sqrt{|T(w_1)| \times |T(w_2)|}}$$

$$\text{sim}_{Dice}(w_1, w_2) = \frac{2 \times |T(w_1) \cap T(w_2)|}{|T(w_1)| + |T(w_2)|}$$

$$\text{sim}_{Jacard}(w_1, w_2) = \frac{|T(w_1) \cap T(w_2)|}{|T(w_1)| + |T(w_2)| - |T(w_1) \cap T(w_2)|}$$

Figure 1: Other Similarity Measures

puted as follows:

$$\begin{aligned} & I(w, r, w') \\ &= -\log(P_{\text{MLE}}(B) P_{\text{MLE}}(A|B) P_{\text{MLE}}(C|B)) \\ &\quad -(-\log P_{\text{MLE}}(A, B, C)) \\ &= \log \frac{||w,r,w|| \times ||*,r,*||}{||w,r,*|| \times ||*,r,w'||} \end{aligned}$$

It is worth noting that $I(w, r, w')$ is equal to the mutual information between $w$ and $w'$ (Hindle, 1990).

Let $T(w)$ be the set of pairs $(r, w')$ such that $\log \frac{||w,r,w'|| \times ||*,r,*||}{||w,r,*|| \times ||*,r,w'||}$ is positive. We define the similarity $\text{sim}(w_1, w_2)$ between two words $w_1$ and $w_2$ as follows:

$$\frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)}$$

We parsed a 64-million-word corpus consisting of the Wall Street Journal (24 million words), San Jose Mercury (21 million words) and AP Newswire (19 million words). From the parsed corpus, we extracted 56.5 million dependency triples (8.7 million unique). In the parsed corpus, there are 5469 nouns, 2173 verbs, and 2632 adjectives/adverbs that occurred at least 100 times. We computed the pairwise similarity between all the nouns, all the verbs and all the adjectives/adverbs, using the above similarity measure. For each word, we created a thesaurus entry which contains the top-N[1] words that are most similar to it.[2] The thesaurus entry for word $w$ has the following format:

$$w\ (pos) : w_1, s_1, w_2, s_2, \ldots, w_N, s_N$$

where $pos$ is a part of speech, $w_i$ is a word, $s_i$=sim$(w, w_i)$ and $s_i$'s are ordered in descending order. For example, the top-10 words in the noun, verb, and adjective entries for the word "brief" are shown below:

**brief (noun):** affidavit 0.13, petition 0.05, memorandum 0.05, motion 0.05, lawsuit 0.05, deposition 0.05, slight 0.05, prospectus 0.04, document 0.04 paper 0.04, ...

**brief (verb):** tell 0.09, urge 0.07, ask 0.07, meet 0.06, appoint 0.06, elect 0.05, name 0.05, empower 0.05, summon 0.05, overrule 0.04, ...

**brief (adjective):** lengthy 0.13, short 0.12, recent 0.09, prolonged 0.09, long 0.09, extended 0.09, daylong 0.08, scheduled 0.08, stormy 0.07, planned 0.06, ...

Two words are a pair of respective nearest neighbors (RNNs) if each is the other's most similar word. Our program found 543 pairs of RNN nouns, 212 pairs of RNN verbs and 382 pairs of RNN adjectives/adverbs in the automatically created thesaurus. Appendix A lists every 10th of the RNNs. The result looks very strong. Few pairs of RNNs in Appendix A have clearly better alternatives.

We also constructed several other thesauri using the same corpus, but with the similarity measures in Figure 1. The measure $\text{sim}_{Hindle}$ is the same as the similarity measure proposed in (Hindle, 1990), except that it does not use dependency triples with negative mutual information. The measure $\text{sim}_{Hindle_r}$ is the same as $\text{sim}_{Hindle}$ except that all types of dependency relationships are used, instead of just subject and object relationships. The measures $\text{sim}_{cosine}$, $\text{sim}_{dice}$ and $\text{sim}_{Jacard}$ are versions of similarity measures commonly used in information retrieval (Frakes and Baeza-Yates, 1992). Unlike sim, $\text{sim}_{Hindle}$ and $\text{sim}_{Hindle_r}$, they only

[1] We used N=200 in our experiments

[2] The resulting thesaurus is available at: http://www.cs.umanitoba.ca/~lindek/sims.htm.

$$\text{sim}_{WN}(w_1, w_2) = \max_{c_1 \in S(w_1) \wedge c_2 \in S(w_2)} (\max_{c \in super(c_1) \cap super(c_2)} \frac{2 \log P(c)}{\log P(c_1) + \log P(c_2)})$$

$$\text{sim}_{Roget}(w_1, w_2) = \frac{2|R(w_1) \cap R(w_2)|}{|R(w_1)| + |R(w_2)|}$$

where $S(w)$ is the set of senses of $w$ in the WordNet, $super(c)$ is the set of (possibly indirect) superclasses of concept $c$ in the WordNet, $R(w)$ is the set of words that belong to a same Roget category as $w$.

Figure 2: Word similarity measures based on WordNet and Roget

make use of the unique dependency triples and ignore their frequency counts.

## 3 Evaluation

In this section, we present an evaluation of automatically constructed thesauri with two manually compiled thesauri, namely, WordNet1.5 (Miller et al., 1990) and Roget Thesaurus. We first define two word similarity measures that are based on the structures of WordNet and Roget (Figure 2). The similarity measure $\text{sim}_{WN}$ is based on the proposal in (Lin, 1997). The similarity measure $\text{sim}_{Roget}$ treats all the words in Roget as features. A word $w$ possesses the feature $f$ if $f$ and $w$ belong to a same Roget category. The similarity between two words is then defined as the cosine coefficient of the two feature vectors.

With $\text{sim}_{WN}$ and $\text{sim}_{Roget}$, we transform WordNet and Roget into the same format as the automatically constructed thesauri in the previous section.

We now discuss how to measure the similarity between two thesaurus entries. Suppose two thesaurus entries for the same word are as follows:

$$w : w_1, s_1, w_2, s_2, \ldots, w_N, s_N$$
$$w : w'_1, s'_1, w'_2, s'_2, \ldots, w'_N, s'_N$$

Their similarity is defined as:

(4)

$$\frac{\sum_{w_i = w'_j} s_i s'_j}{\sqrt{(\sum_{i=1}^{N} s_i^2)(\sum_{j=1}^{N} s'^2_j)}}$$

For example, (5) is the entry for "brief (noun)" in our automatically generated thesaurus and (6) and (7) are corresponding entries in WordNet thesaurus and Roget thesaurus.

(5) brief (noun): affidavit 0.13, petition 0.05, memorandum 0.05, motion 0.05, lawsuit 0.05, deposition 0.05, slight 0.05, prospectus 0.04, document 0.04 paper 0.04.

(6) brief (noun): outline 0.96, instrument 0.84, summary 0.84, affidavit 0.80, deposition 0.80, law 0.77, survey 0.74, sketch 0.74, resume 0.74, argument 0.74.

(7) brief (noun): recital 0.77, saga 0.77, autobiography 0.77, anecdote 0.77, novel 0.77, novelist 0.77, tradition 0.70, historian 0.70, tale 0.64.

According to (4), the similarity between (5) and (6) is 0.297, whereas the similarities between (5) and (7) and between (6) and (7) are 0.

Our evaluation was conducted with 4294 nouns that occurred at least 100 times in the parsed corpus and are found in both WordNet1.5 and the Roget Thesaurus. Table 1 shows the average similarity between corresponding entries in different thesauri and the standard deviation of the average, which is the standard deviation of the data items divided by the square root of the number of data items. Since the differences among $\text{sim}_{cosine}$, $\text{sim}_{dice}$ and $\text{sim}_{Jacard}$ are very small, we only included the results for $\text{sim}_{cosine}$ in Table 1 for the sake of brevity.

It can be seen that sim, $\text{Hindle}_r$ and cosine are significantly more similar to WordNet than Roget is, but are significantly less similar to Roget than WordNet is. The differences between Hindle and $\text{Hindle}_r$ clearly demonstrate that the use of other types of dependencies in addition to subject and object relationships is very beneficial.

The performance of sim, $\text{Hindle}_r$ and cosine are quite close. To determine whether or not the differences are statistically significant, we computed their differences in similarities to WordNet and Roget thesaurus for each individual entry. Table 2 shows the average and standard deviation of the average difference. Since the 95% confidence inter-

Table 1: Evaluation with WordNet and Roget

| | WordNet | |
|---|---|---|
| | average | $\sigma_{avg}$ |
| Roget | 0.178397 | 0.001636 |
| sim | 0.212199 | 0.001484 |
| $\text{Hindle}_r$ | 0.204179 | 0.001424 |
| cosine | 0.199402 | 0.001352 |
| Hindle | 0.164716 | 0.001200 |

| | Roget | |
|---|---|---|
| | average | $\sigma_{avg}$ |
| WordNet | 0.178397 | 0.001636 |
| sim | 0.149045 | 0.001429 |
| $\text{Hindle}_r$ | 0.14663 | 0.001383 |
| cosine | 0.135697 | 0.001275 |
| Hindle | 0.115489 | 0.001140 |

vals of all the differences in Table 2 are on the positive side, one can draw the statistical conclusion that simis better than $\text{sim}_{Hindle_r}$, which is better than $\text{sim}_{cosine}$.

Table 2: Distribution of Differences

| | WordNet | |
|---|---|---|
| | average | $\sigma_{avg}$ |
| $\text{sim}-\text{Hindle}_r$ | 0.008021 | 0.000428 |
| sim−cosine | 0.012798 | 0.000386 |
| $\text{Hindle}_r-\text{cosine}$ | 0.004777 | 0.000561 |

| | Roget | |
|---|---|---|
| | average | $\sigma_{avg}$ |
| $\text{sim}-\text{Hindle}_r$ | 0.002415 | 0.000401 |
| sim−cosine | 0.013349 | 0.000375 |
| $\text{Hindle}_r-\text{cosine}$ | 0.010933 | 0.000509 |

## 4 Future Work

Reliable extraction of similar words from text corpus opens up many possibilities for future work. For example, one can go a step further by constructing a tree structure among the most similar words so that different senses of a given word can be identified with different subtrees. Let $w_1, \ldots, w_n$ be a list of words in descending order of their similarity to a given word $w$. The similarity tree for $w$ is created as follows:

- Initialize the similarity tree to consist of a single node $w$.
- For $i$=1, 2, ..., $n$, insert $w_i$ as a child of $w_j$ such that $w_j$ is the most similar one to $w_i$ among $\{w, w_1, \ldots, w_{i-1}\}$.

For example, Figure 3 shows the similarity tree for the top-40 most similar words to duty. The first number behind a word is the similarity of the word to its parent. The second number is the similarity of the word to the root node of the tree.

```
duty
|___responsibility 0.21 0.21
|   |___role 0.12 0.11
|   |   |___action 0.11 0.10
|   |   |   |___change 0.24 0.08
|   |   |   |   |___rule 0.16 0.08
|   |   |   |   |   |___restriction 0.27 0.08
|   |   |   |   |   |   |___ban 0.30 0.08
|   |   |   |   |   |   |___sanction 0.19 0.08
|   |   |   |   |   |___schedule 0.11 0.07
|   |   |   |   |   |___regulation 0.37 0.07
|   |   |   |___challenge 0.13 0.07
|   |   |   |   |___issue 0.13 0.07
|   |   |   |   |   |___reason 0.14 0.07
|   |   |   |   |   |___matter 0.28 0.07
|   |   |   |___measure 0.22 0.07
|   |___obligation 0.12 0.10
|   |___power 0.17 0.08
|   |   |___jurisdiction 0.13 0.08
|   |   |___right 0.12 0.07
|   |   |___control 0.20 0.07
|   |   |___ground 0.08 0.07
|   |___accountability 0.14 0.08
|   |___experience 0.12 0.07
|___post 0.14 0.14
|   |___job 0.17 0.10
|   |   |___work 0.17 0.10
|   |   |   |___training 0.11 0.07
|   |___position 0.25 0.10
|___task 0.10 0.10
|   |___chore 0.11 0.07
|___operation 0.10 0.10
|   |___function 0.10 0.08
|   |___mission 0.12 0.07
|   |   |___patrol 0.07 0.07
|   |___staff 0.10 0.07
|___penalty 0.09 0.09
|   |___fee 0.17 0.08
|   |   |___tariff 0.13 0.08
|   |   |___tax 0.19 0.07
|___reservist 0.07 0.07
```

Figure 3: Similarity tree for "duty"

Inspection of sample outputs shows that this algorithm works well. However, formal evaluation of its accuracy remains to be future work.

## 5 Related Work and Conclusion

There have been many approaches to automatic detection of similar words from text corpora. Ours is

similar to (Grefenstette, 1994; Hindle, 1990; Ruge, 1992) in the use of dependency relationship as the word features, based on which word similarities are computed.

Evaluation of automatically generated lexical resources is a difficult problem. In (Hindle, 1990), a small set of sample results are presented. In (Smadja, 1993), automatically extracted collocations are judged by a lexicographer. In (Dagan et al., 1993) and (Pereira et al., 1993), clusters of similar words are evaluated by how well they are able to recover data items that are removed from the input corpus one at a time. In (Alshawi and Carter, 1994), the collocations and their associated scores were evaluated indirectly by their use in parse tree selection. The merits of different measures for association strength are judged by the differences they make in the precision and the recall of the parser outputs.

The main contribution of this paper is a new evaluation methodology for automatically constructed thesaurus. While previous methods rely on indirect tasks or subjective judgments, our method allows direct and objective comparison between automatically and manually constructed thesauri. The results show that our automatically created thesaurus is significantly closer to WordNet than Roget Thesaurus is. Our experiments also surpasses previous experiments on automatic thesaurus construction in scale and (possibly) accuracy.

## Acknowledgement

This research has also been partially supported by NSERC Research Grant OGP121338 and by the Institute for Robotics and Intelligent Systems.

## References

Hiyan Alshawi and David Carter. 1994. Training and scaling preference functions for disambiguation. *Computational Linguistics*, 20(4):635–648, December.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proceedings of ACL-93*, pages 164–171, Columbus, Ohio, June.

Ido Dagan, Fernando Pereira, and Lillian Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd Annual Meeting of the ACL*, pages 272–278, Las Cruces, NM.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based method for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the ACL*, pages 56–63, Madrid, Spain.

Ute Essen and Volker Steinbiss. 1992. Cooccurrence smoothing for stochastic language modeling. In *Proceedings of ICASSP*, volume 1, pages 161–164.

W. B. Frakes and R. Baeza-Yates, editors. 1992. *Information Retrieval, Data Structure and Algorithms*. Prentice Hall.

D. Gentner. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj, editor, *Language development: Vol. 2. Language, thought, and culture*, pages 301–334. Erlbaum, Hillsdale, NJ.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston, MA.

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of ACL-90*, pages 268–275, Pittsburg, Pennsylvania, June.

Dekang Lin. 1993. Principle-based parsing without overgeneration. In *Proceedings of ACL-93*, pages 112–120, Columbus, Ohio.

Dekang Lin. 1994. Principar—an efficient, broad-coverage, principle-based parser. In *Proceedings of COLING-94*, pages 482–488. Kyoto, Japan.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACL/EACL-97*, pages 64–71, Madrid, Spain, July.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

George A. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

Eugene A. Nida. 1975. *Componential Analysis of Meaning*. The Hague, Mouton.

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional Clustering of English Words. In *Proceedings of ACL-93*, pages 183–190, Ohio State University, Columbus, Ohio.

Gerda Ruge. 1992. Experiments on linguistically based term associations. *Information Processing & Management*, 28(3):317–332.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–178.

## Appendix A: Respective Nearest Neighbors

| Nouns | | |
|---|---|---|
| Rank | Respective Nearest Neighbors | Similarity |
| 1 | earnings profit | 0.572525 |
| 11 | plan proposal | 0.47475 |
| 21 | employee worker | 0.413936 |
| 31 | battle fight | 0.389776 |
| 41 | airline carrier | 0.370589 |
| 51 | share stock | 0.351294 |
| 61 | rumor speculation | 0.327266 |
| 71 | outlay spending | 0.320535 |
| 81 | accident incident | 0.310121 |
| 91 | facility plant | 0.284845 |
| 101 | charge count | 0.278339 |
| 111 | baby infant | 0.268093 |
| 121 | actor actress | 0.255098 |
| 131 | chance likelihood | 0.248942 |
| 141 | catastrophe disaster | 0.241986 |
| 151 | fine penalty | 0.237606 |
| 161 | legislature parliament | 0.231528 |
| 171 | oil petroleum | 0.227277 |
| 181 | strength weakness | 0.218027 |
| 191 | radio television | 0.215043 |
| 201 | coupe sedan | 0.209631 |
| 211 | turmoil upheaval | 0.205841 |
| 221 | music song | 0.202102 |
| 231 | bomb grenade | 0.198707 |
| 241 | gallery museum | 0.194591 |
| 251 | leaf leave | 0.192483 |
| 261 | fuel gasoline | 0.186045 |
| 271 | door window | 0.181301 |
| 281 | emigration immigration | 0.176331 |
| 291 | espionage treason | 0.17262 |
| 301 | peril pitfall | 0.169587 |
| 311 | surcharge surtax | 0.166831 |
| 321 | ability credibility | 0.163301 |
| 331 | pub tavern | 0.158815 |
| 341 | license permit | 0.156963 |
| 351 | excerpt transcript | 0.150941 |
| 361 | dictatorship regime | 0.148837 |
| 371 | lake river | 0.145586 |
| 381 | disc disk | 0.142733 |
| 391 | interpreter translator | 0.138778 |
| 401 | bacteria organism | 0.135539 |
| 411 | ballet symphony | 0.131688 |
| 421 | silk wool | 0.128999 |
| 431 | intent intention | 0.125236 |
| 441 | waiter waitress | 0.122373 |
| 451 | blood urine | 0.118063 |
| 461 | mosquito tick | 0.115499 |
| 471 | fervor zeal | 0.112087 |
| 481 | equal equivalent | 0.107159 |
| 491 | freezer refrigerator | 0.103777 |
| 501 | humor wit | 0.0991108 |
| 511 | cushion pillow | 0.0944567 |
| 521 | purse wallet | 0.0914273 |
| 531 | learning listening | 0.0859118 |
| 541 | clown cowboy | 0.0714762 |

| Verbs | | |
|---|---|---|
| Rank | Respective Nearest Neighbors | Similarity |
| 1 | fall rise | 0.674113 |
| 11 | injure kill | 0.378254 |
| 21 | concern worry | 0.340122 |
| 31 | convict sentence | 0.289678 |
| 41 | limit restrict | 0.271588 |
| 51 | narrow widen | 0.258385 |
| 61 | attract draw | 0.242331 |
| 71 | discourage encourage | 0.234425 |
| 81 | hit strike | 0.22171 |
| 91 | disregard ignore | 0.21027 |
| 101 | overstate understate | 0.199197 |
| 111 | affirm reaffirm | 0.182765 |
| 121 | inform notify | 0.170477 |
| 131 | differ vary | 0.161821 |
| 141 | scream yell | 0.150168 |
| 151 | laugh smile | 0.142951 |
| 161 | compete cope | 0.135869 |
| 171 | add whisk | 0.129205 |
| 181 | blossom mature | 0.123351 |
| 191 | smell taste | 0.112418 |
| 201 | bark howl | 0.101566 |
| 211 | black white | 0.0694954 |

| Adjective/Adverbs | | |
|---|---|---|
| Rank | Respective Nearest Neighbors | Similarity |
| 1 | high low | 0.580408 |
| 11 | bad good | 0.376744 |
| 21 | extremely very | 0.357606 |
| 31 | deteriorating improving | 0.332664 |
| 41 | alleged suspected | 0.317163 |
| 51 | clerical salaried | 0.305448 |
| 61 | often sometimes | 0.281444 |
| 71 | bleak gloomy | 0.275557 |
| 81 | adequate inadequate | 0.263136 |
| 91 | affiliated merged | 0.257666 |
| 101 | stormy turbulent | 0.252846 |
| 111 | paramilitary uniformed | 0.246638 |
| 121 | sharp steep | 0.240788 |
| 131 | communist leftist | 0.232518 |
| 141 | indoor outdoor | 0.224183 |
| 151 | changed changing | 0.219697 |
| 161 | defensive offensive | 0.211062 |
| 171 | sad tragic | 0.206688 |
| 181 | enormously tremendously | 0.199936 |
| 191 | defective faulty | 0.193863 |
| 201 | concerned worried | 0.186899 |
| 211 | dropped fell | 0.184768 |
| 221 | bloody violent | 0.183058 |
| 231 | favorite popular | 0.179234 |
| 241 | permanently temporarily | 0.174361 |
| 251 | confidential secret | 0.17022 |
| 261 | privately publicly | 0.165313 |
| 271 | operating sales | 0.162894 |
| 281 | annually apiece | 0.159883 |
| 291 | gentle kind | 0.154554 |
| 301 | losing winning | 0.149447 |
| 311 | experimental test | 0.146435 |
| 321 | designer dress | 0.142552 |
| 331 | dormant inactive | 0.137002 |
| 341 | commercially domestically | 0.132918 |
| 351 | complimentary free | 0.128117 |
| 361 | constantly continually | 0.122342 |
| 371 | hardy resistant | 0.112133 |
| 381 | anymore anyway | 0.103241 |