Modeling Information Scent: A Comparison of LSA, PMI and GLSA Similarity Measures on Common Tests and Corpora

Raluca Budiu, Christiaan Royer and Peter Pirolli

Palo Alto Research Center, Palo Alto, CA 94087

{budiu,royer,pirolli}@parc.com

Abstract

In this paper we describe a comparison among three systems that estimate semantic similarity between words: Latent Semantic Analysis (Landauer & Dumais, 1997), Pointwise Mutual Information (Turney, 2001), and Generalized Latent Semantic Analysis (Matveeva, Levow, Farahat, & Royer, 2005). We compare all these techniques on a unique corpus (TASA) and, for PMI and GLSA, we also report performance on a larger web-based corpus. The evaluation is carried out through two kinds of tests: (1) synonymy tests, and (2) comparison with human word similarity judgments. The results indicate that for large corpora PMI works best on word similarity tests, and GLSA on synonymy tests. For the smaller TASA corpus, GLSA produced the best performance on most tests. A large corpus improved the performance of PMI, but, in most cases, did not improve that of GLSA.

Introduction

A problem often encountered when browsing for information in large document collections such as the World Wide Web is finding relevant information: when searching for a particular target, users must continually assess the semantics of labeled navigation options (e.g., Web links) and judge the relevance and utility of those options. User studies have shown that navigation choices are largely driven by how well the link labels semantically match the search goal. The semantic cues that the navigation choices offer have been known as **information scent** (Pirolli & Card, 1999). Information scent is now a central construct in cognitive models of users (Pirolli, 2005), usability testing systems (Chi, Rosien, & Suppattanasiri, 2003), and design guidelines (Spool, Perfetti, & Brittan, 2004).

Information scent is perhaps most obvious in the case of browsing a web site for particular information. We often decide what links to follow based on our knowledge

about the desired target information and on our intuitions that a particular set of link label words may be relevant to that target information. Theoretical analyses (Pirolli, 2005) and empirical usability research (Spool et al., 2004) suggest that information scent is the most important factor in Web navigation. In the field of information visualization (Card, Mackinlay, & Schneiderman, 1999), research (Pirolli, Card, & Van Der Wege, 2000; Budiu & Pirolli, 2006) shows that, although the particular information visualization often has little effect on user performance, what impacts the user performance most is information scent.

However, a question that arises with the notion of information scent is how do we measure it? Often, it is difficult to model the scent of the information that is displayed on the screen: if the users are looking for *peanut butter*, will they click on the word *jelly*, or on the word *nuts* or perhaps *tropical fruits*? A good measure of information scent would presumably achieve two goals: (1) supply engineering techniques for designing better interfaces and web pages that display "good information scent" and lead people on the right search paths; (2) help us design smarter information retrieval systems that would better guess what people have in mind when they search for a target, by comparing information scent for candidate documents.

One possible measure of information scent is semantic similarity. However, semantic similarity per se is hard to measure simply because it is not feasible to ask people for similarity ratings of all words in a language that may be used in an user interface or on the Web. Fortunately, several techniques that automatically estimate word similarity have emerged. Many of these techniques are based on word co-occurrence in a large corpus: two words are similar depending on how often they co-occur or on how often they occur in the same context. One problem for the modeler is which of these techniques to use? Is there one that is better than another? Attempts to compare these techniques have been made in the past (Kaur & Hornof, 2005; Rohde, Gonnerman, & Plaut, 2006; Terra & Clarke, 2003; Turney, 2001). Such comparisons typically involve two kinds of tasks: (1) synonymy tests, in which the systems have to choose the word most similar to a target word; and (2) similarity tests, in which the systems provide their estimate of the similarity between two words; then those numbers are compared with the similarity ratings collected from people.

Unfortunately, any co-occurrence-based estimate of similarity is bound to be dependent on the corpus on which those co-occurrences were computed. Thus, if the corpus contained only Computer Science documents, probably the words *apple* and *computer* would occur in the same contexts often, and thus would be highly similar. However, most people would rate these words as far apart. We believe that one major deficiency of the previous studies has been using these different estimates of similarity on different corpora. Thus, if a study deems a method superior to another, it is not clear how much that is the merit of the method per se, and how much it is the merit of the corpus. In this pa-

per we examine how three such techniques: Latent Semantic Analysis (LSA)(Landauer & Dumais, 1997), Pointwise Mutual Information (PMI) (Turney, 2001), and Generalized Latent Semantic Analysis (GLSA) (Matveeva et al., 2005) compare with each other, when co-occurrence counts are based on the same corpus. The common corpus is that used for creating the word frequency guide published by Touchstone Applied Science Associates (TASA) (Zeno, Ivens, Millard, & Duvvuri, 1995). It is one of the corpora available at the official LSA web interface. Albeit its many advantages, TASA is not a public domain corpus and will become inherently dated. One of the hopes nurtured by many in the information retrieval community is that the large amount of text available on the web could be used as a corpus. Then, the other question that we address in this paper is how the web (or a large sample of it) compares with a carefully constructed corpus such as TASA. Although, for reasons explained below, LSA could not be run on a large web-based corpus, we discuss how the TASA corpus and a larger web-based corpus compare to each other when they are used as base corpora for PMI and GLSA.

In the rest of the paper we describe the three techniques that we compare, the corpora on which we run the tests, and the evaluation results. We end with reviewing other related work and with conclusions.

Latent Semantic Analysis (LSA)

LSA (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998) is a method for computing semantic distance between words and has been repeatedly used as a measure of semantic similarity. It is based on a collection of different documents. LSA represents words as vectors in a word-by-document space; a summary of this method is given in Table 1. Given a corpus of text that can be split into documents (some being just paragraphs in a larger text), for each word, LSA computes how often that word occurs in each document (step 1). The matrix C of word-by-document occurrences is then normalized (steps 2-4), and then singular value decomposition is applied to obtain a lower dimensionality approximation of the original matrix. The rank of the new matrix is k, the number of factors used to perform the decomposition. Words are represented as vectors in this new matrix; the similarity between two words is estimated as the cosine of the angle between the two corresponding vectors.

Here we report LSA values obtained from the LSA website: http://lsa.colorado.edu. The website lets users specify different semantic spaces (i.e., corpora) for the similarity computation; for all our computations we use the semantic space labeled as *General Reading up to 1st year college*, which corresponds to the TASA corpus (see below). Because LSA uses a word-by-document matrix, the cost of computation increases substantially with large semantic spaces.

According to (Landauer & Dumais, 1997; Landauer et al., 1998), LSA was able to

For a set of words W and a set of documents D and a number of factors k:

1. Compute the matrix C of word-by-document occurrences: $C[i, j]$ represents how many times word i occurs in document j.
2. Compute LC from C such that $LC[i, j] = \log(1 + C[i, j])$.
3. Compute the entropy $H[i]$ of word i as
$$H[i] = \sum_j -C[i, j] log C[i, j].$$
4. Normalize the entries in LC: $N[i, j] = LC[i, j]/H[i]$.
5. Use singular value decomposition on LC to obtain a matrix Q of dimensionality k.
6. A word i is represented as the vector $Q[i]$ and the similarity between words i and j is $\cos(Q[i], Q[j])$.

Table 1: A description of LSA.

score 64.4% in the TOEFL synonymy test. LSA was also used to simulate behavioral data from psycholinguistics experiments (e.g., subject-matter knowledge, semantic priming effects, metaphor, learning from text) and was also successfully applied to domains as varied as essay scoring, predicting learning from text (see Landauer et al., 1998), and web page evaluation (Blackmon, Kitajima, & Polson, 2005).

Pointwise Mutual Information (PMI)

The Pointwise Mutual Information (PMI) (e.g., Manning & Schutze, 1999) between two words A and B captures how likely it is to find B in a text given that you know that the text contains A. It is a co-occurrence metric, in that it normalizes the probability of co-occurrence of the two words with their individual probabilities of co-occurrence. The PMI between A and B can be calculated as:

$$PMI(A, B) = \log \frac{p(A, B)}{p(A)p(B)} \approx \log \frac{C(A, B) \times N}{C(A)C(B)}$$

where $p(A, B)$ is the probability that A and B co-occur in the same document (possibly within a text window of a fixed size); $p(X)$ is the probability that word X occurs in a document; $C(A, B)$ is the number of documents in which A and B co-occur; $C(X)$ is the number of documents in which X occurs, and N is the number of documents in the corpus. For our intents and purposes N can be ignored (since it is the same value in all PMI pairs and since we are interested in the ordering of PMI scores). The logarithm can also be removed, since we are interested in a relative ordering of pairs. Then the PMI can be approximated as:

$$PMI(A, B) \approx \frac{C(A, B)}{C(A)C(B)}$$

The similarity between words A and B is then estimated by their PMI score.

In our tests we use a window of 16 words to capture co-occurrence. A web interface for computing PMI is available at http://glsa.parc.com. Turney (2001) used PMI on the corpus from the Alta Vista search engine to estimate word similarities. He used synonymy tests to compare the results obtained from PMI and those obtained from LSA, run on a different smaller corpus. PMI resulted in performance comparable or better (between 62.5% and 73.75%) than LSA on TOEFL.

Generalized Latent Semantic Analysis (GLSA)

GLSA (Matveeva et al., 2005) is a technique that combines the strengths of both PMI and LSA. Like LSA, it uses dimensionality reduction (SVD) to filter out noise in the system. Unlike LSA, the initial word-by-document co-occurrence matrix is replaced by a word-by-word PMI matrix, in which words are represented as vectors of PMI scores relative to other words in the vocabulary (Niwa & Nitta, 1994).

Table 2 presents the basic steps for GLSA. GLSA starts with a large document corpus C and a subset of words V (called **terms**). Then (steps 1 and 2) it computes a word-by-word matrix in which each entry represents the PMI (as computed above, within a 16-word window) between words i and j in the vocabulary V. Then, as for LSA, a SVD is applied to the resulting matrix and the similarity between the two words is the cosine of the corresponding vectors in the reduced matrix. In the case of GLSA there are two parameters: (1) the number of factors k, representing the rank to which the matrix is reduced (similar to LSA) ; and (2) the number of terms to be included in the vocabulary V.

Unlike LSA, because GLSA only computes a word-by-word matrix corresponding to a restricted vocabulary V, the cost of GLSA computations does not depend as much on

Given a set of words V, a large document corpus D, and a number of factors k:

1. Compute the co-occurrence matrix C: $C[i, j]$ represents how often words i and j in vocabulary V co-occur in the corpus D within a $n = 16$ word window; for each word i also compute $F[i]$: how often it occurs in the corpus
2. Compute the PMI matrix: $PMI[i, j] = \log \frac{C[i,j]}{F(i)F(j)}$.
3. Perform singular value decomposition on the PMI matrix to reduce the dimensionality of PMI to a lower rank k; the resulting matrix is Q.
4. Compute the similarity between words i and j as the cosine of their corresponding vectors in the matrix Q: $\cos Q[i]Q[j]$.

Table 2: A summary of GLSA.

the size of the corpus[1]. However, only similarities corresponding to words from the term list V can be computed.

In the experiment reported in (Matveeva et al., 2005), GLSA obtained 86% performance on TOEFL, using a corpus made of New York Times articles. A web interface for GLSA is available at http://glsa.parc.com.

Corpora

We evaluate PMI, LSA and GLSA on two corpora: the TASA corpus and a web-based corpus called Stanford corpus.

The TASA Corpus

The TASA corpus was created in 1995 from "60,527 samples of text from 6,333 textbooks, works of literature, and popular works of fiction and nonfiction used in schools and colleges throughout the United States." The main purpose behind this corpus was to estimate the frequency of words encountered by school-age American children at differ-

[1]The computational cost of GLSA does depend on the size of the corpus because for each term co-occurrence counts must be computed; however, the matrix C is much smaller than the corresponding matrix in LSA.

ent grade levels. Thus, the corpus represents much of the textual material that typical US students are likely to encounter in their school years and it contains a representative sample of the texts that students at a certain grade level are expected to read. It includes textbooks, texts from recommended or required reading lists, and materials submitted by states, school districts, or colleges to TASA for analyses of text difficulty or readability. The TASA corpus covers nine content areas (language arts and literature, social science, science and math, fine arts, home economics and related fields, trade, service and technical fields, health, safety and related fields, business and related fields, popular fiction and nonfiction). Most of the texts (85%) included in this corpus were published between 1971 and 1995.

In building the corpus, the authors followed a systematic plan for sampling the texts, based on the length of the book. The book was divided in sections of equal length and several samples were taken from each section. The samples were around 250-325 words.

The final TASA corpus contains 17,274,580 tokens[2] corresponding to 154,941 different types[3]. TASA subdivided the corpus in different corpora corresponding to grades K up to college; however, for our analyses, we use the entire undivided corpus.

On the LSA website, the TASA corpus is labeled as *General reading up to 1st year college*.

The Stanford Corpus

The second corpus used for our tests was gathered by web crawling by the Stanford WebBase project (`http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/`). Unlike the TASA corpus, which was very carefully put together, the WebBase corpus aims to be a local significant reflection of the Web (Cho et al., 2004). WebBase is an academic continuation of the original Google repository and search engine. The crawler is seeded with a number of initial sites, and then new sites linked from the initial sites are added to the repository.

Throughout the paper we refer to the first 6.7 million pages from this project as the Stanford corpus. The Stanford corpus contains approximately 30 million different types. Due to the high number of documents in this corpus and to the fact that LSA computes a word by document matrix, it was not possible to run LSA on the Stanford corpus. We report only PMI and GLSA results on that corpus. However, we did run all of LSA, PMI and GLSA on the TASA corpus, so we can fairly compare them at least on one corpus.

[2]The term **token** refers to the total number of words in the corpus.

[3]The term **type** refers to different appearances of the same word: for instance, if the word *apple* occurred 32 times in a corpus, it would contribute only once to the number of types, but 32 times to the number of tokens.

Evaluation

We are interested in finding how well our three measures of choice (LSA, PMI, and GLSA) approximate semantic similarity as understood by people. To do that, we compare the results provided by these metrics with those obtained from people on two kinds of tests: (1) synonymy tests, and (2) similarity ratings.

For LSA, all the results are obtained using the LSA web interface. For each test set, we varied the number of factors from 10 to 300 (with an increment of 10). (Note that 300 is the maximum number of factors allowed by the interface.)

For GLSA and PMI we only counted word co-occurrences within a 16-word window. Turney (2001) has pointed out that window methods work better than counting how many documents contain both words. The number of factors for GLSA was varied between 10 and 2000. The results from GLSA were obtained using a vocabulary of approximately 33,000 terms. These terms included all the words in the tests. One of our tests (Reader's Digest) included multiple-word phrases, so we created two variants of the term lists: one including each phrase as a term (GLSA-MW condition), and the second including each individual word in the phrase as a term (GLSA-SW condition). The term list for GLSA-MW contained 32,714 words, whereas the one for GLSA-SW contained 32,554 terms.

Table 3 summarizes the results of using the three methods on a variety of benchmarks. The table reports the best result obtained by each method, when the number of factors was varied as described previously. In what follows we describe these benchmarks in detail.

Synonymy Tests

Synonymy tests are often used for testing vocabulary knowledge. A question in a synonymy test presents a word and asks the test taker to choose another word most similar to it among several (usually four) alternatives. The performance on this test is computed as the percentage of questions answered correctly out of the total number of questions. We use three synonymy tests: Test of English as a Foreign Language (TOEFL), first used in (Landauer & Dumais, 1997), English as a Second Language (ESL) (Turney, 2001), and Reader's Digest Word Power Vocabulary Test (RD). TOEFL and ESL are intended for foreign students.

TOEFL. TOEFL has 80 questions that are drawn from testing materials designed by the Educational Testing Service primarily to measure the knowledge of English of foreign students coming to the US. TOEFL has been first used by Landauer and Dumais (1997) to compare the results obtained by LSA with those of students. The items in the 80 questions are all single words; a sample question is to select the synonym of the words *flawed* from a set containing *imperfect*, *tiny*, *lustrous*, and *crude*. Landauer and Dumais

	TASA				Stanford		
	LSA	PMI	GLSA-MW	GLSA-SW	PMI	GLSA-MW	GLSA-SW
TOEFL	60 (140)	23	**72** (72)	71 (44)	51	73 (254)	**76** (249)
ESL	44 (300)	22	59 (63)	**60**(624)	52	58 (679)	**60** (220)
RD	46(300)	22	52 (15)	**55** (27)	40	66(332)	**67** (514)
WS353	60 (280)	58	**65** (108)	**65** (82)	**71**	54 (158)	55 (159)
MC	75 (250)	65	**79** (33)	77(103)	**79**	61(139)	62 (137)
R	71 (220)	61	79 (33)	**81** (61)	**83**	60 (140)	61 (139)
RG	64 (250)	61	**76** (70)	75(64)	**75**	68 (164)	69 (155)
RGP	**63** (300)	47	56 (54)	56 (53)	51	**55** (118)	**55** (325)
N	**25** (290)	21	21 (413)	22(472)	14	19(1703)	**20** (1664)

Table 3: Performance of LSA, PMI and GLSA-MW, GLSA-SW: percent correct on synonymy tests (TOEFL, ESL, RD) and rank correlation (*100) on word similarity tests (WS353, MC, R, RG, RGP, N). The corpora used (TASA and Stanford). The number in parentheses in each cell (for LSA and GLSA) represents the number of factors used to obtain the best performance.

(1997) report that foreign college students typically score around 65% on the synonymy questions in TOEFL. On average, a word in this test occurred in about 56,656 documents in the Stanford corpus and 423 documents in the TASA corpus.

Our results indicate that GLSA-MW was the best method for the TASA corpus, achieving a maximum accuracy of 72%, higher than both PMI and LSA. The difference between the two term lists for GLSA (GLSA-MW and GLSA-SW) was minimal. The value we obtained for LSA was lower than the 65% value reported by Landauer et al. (1998). (More recent changes in the LSA algorithm or corpus may have determined this difference.) GLSA was also superior to PMI on the Stanford corpus; GLSA-SW did slightly better than GLSA-MW. Unlike others (Turney, 2001; Terra & Clarke, 2003), we did not find very good performance for PMI on either TASA or Stanford. PMI seems to work very well with large corpora, and Stanford and TASA may have been still too small for this technique, at least as far as TOEFL is concerned.

ESL. ESL has 57 questions, containing single words. ESL was first used by Turney (2001) to compare the performance of PMI and LSA. Sample items include *rusty* (choices: *corroded, black, dirty, painted*) and *lump* (choices: *chunk, stem, trunk, limb*). On average, the words in the ESL test occurred in 501 documents in TASA and 71,559 in Stanford; thus, the words in this test were more frequent than in TOEFL.

On TASA corpus, the results show that GLSA was best on this test, with an accuracy of 60% for GLSA-SW 59% for GLSA-MW, followed by LSA and PMI. Using the

bigger Stanford corpus did not improve much the accuracy of GLSA, but it substantially improved the performance of PMI. Still, on the Stanford corpus GLSA performed best (60% and 58% for GLSA-SW, respectively GLSA-MW).

RD. The Reader's Digest Word Power Test contains 103 questions from Reader's Digest word quizzes. It has been first used by Jarmasz and Szpakowicz (2003). Some of the items involved are multiple-word phrases (e.g., *insipid* has choices *not interesting*, *slow moving*, *lacking in thoroughness*, and *easygoing*). Each word or multiple-word phrase in the RD test occurred on average in 24,773 documents in the Stanford corpus and in 130 documents from the TASA corpus.

For LSA we computed the performance by introducing the multiple-word phrases in the interface. LSA calculates semantic distance among different texts by first summing the vectors corresponding to the words in each passage and obtaining two document vectors. Then the cosine between the two document vectors estimates the similarity between the corresponding documents. For GLSA, we either included the phrases in the term list (GLSA-MW) or we used component words as terms (GLSA-SW). For GLSA-SW we used the same method as LSA to compute the phrase-to-phrase similarity. In GLSA-MW, the phrase was treated as a regular word and corresponded to a row in the reduced-dimensionality matrix.

For PMI we counted the co-occurrences of those exact word phrases; the exact word phrases were also included in the term list for GLSA.

GLSA-SW was best on this test, achieving 55% correct on TASA and 67% correct on Stanford. (GLSA-MW was also very close.) The performance of PMI was low on TASA, probably due to the small corpus; using a larger corpus increased performance by about 20 percentage points. The larger corpus also helped GLSA. Interestingly, the two term lists for GLSA (one containing the multiple-word phrases encountered specifically in this test) did not lead to a substantial difference in performance, supporting the idea of decomposing phrases into component words and using words as terms, then deriving phrase similarity based on vectors of component words.

Overall, for the synonymy tests, the best technique proved to be GLSA. The larger corpus tended to improve performance, although less for GLSA than for PMI. Moreover, GLSA-MW and GLSA-SW rendered very close results on synonymy tests.

Similarity Ratings

Another way of measuring the performance of a similarity metric is to compare with similarity ratings collected from people. Given a list of word pairs and associated human ratings, for each word pair we first average all the ratings obtained from all raters. Then

we compute a rank correlation between the estimates and the average ratings. We use several existing data sets: (1) Rubenstein and Goodenough (RG) ratings (Rubenstein & Goodenough, 1965); (2) Miller and Charles (MC) ratings (Miller & Charles, 1991); (3) Resnick (R) ratings (Resnick, 1995); (4) the Word Similarity Test Collection (WS353) (Finkelstein et al., 2002); (5) Rohde, Gonnerman and Plaut's (RGP) ratings (RGP) (Rohde et al., 2006); (6) Nelson (N)'s ratings (Nelson, Dyrdal, & Goodmon, 2005).

RG (Rubenstein & Goodenough, 1965). This set of ratings contains 65 noun pairs, rated on a scale from 0 to 4 by 51 human raters. It was collected back in 1965. Sample pairs include *asylum – cemetery* and *monk – oracle*. The words in this test were present, on average, in 354 documents from TASA and 29938 documents from the Stanford corpus.

The performance on this test was best for GLSA-MW and GLSA-SW (76% and respectively 75% rank correlation), when we used the TASA corpus, with LSA and PMI giving results relatively close to each other. However, on the Stanford corpus, PMI scored the best (75%); again the two variants of GLSA differed very little. Interestingly, using the Stanford corpus deteriorated the performance of GLSA.

MC (Miller & Charles, 1991). This is a 30-pair subset of the list used by Rubenstein and Goodenough (1965); the ratings were provided by 38 subjects. This subset correlated very well (0.97) to the original ratings collected by Rubenstein and Goodenough (1965). The words in this test occurred in 512 TASA documents (on average) and in 42,435 texts from Stanford.

The best performance belonged to GLSA on the TASA corpus (79% for GLSA-MW and 77% for GLSA-SW). LSA was close behind at 75%. On the Stanford corpus, PMI did best. Again, using the Stanford corpus deteriorated the performance of GLSA.

R (Resnick, 1995). Resnick replicated in 1995 the ratings for the MC set, using 10 human raters. The mean ratings in his dataset correlate at 0.96 level with the MC set.

GLSA-SW correlates best with users (81%) when using the TASA corpus, followed very closely by GLSA-MW. On Stanford, PMI is much better. Again, GLSA performs best on the smaller corpus, whereas PMI is helped by the larger size corpus.

WS353 (Finkelstein et al., 2002). Finkelstein et al. (2002) collected ratings from 13 or 16 subjects for 353 word pairs, on a scale from 0 to 10. The MC set was also included in their subset; the correlation between their ratings and those collected by (Miller & Charles, 1991) was 0.95. This dataset also contained nouns, but also proper nouns, adjectives, and gerunds. Examples of pairs include *smart – stupid, smart – student, FBI – fingerprint*. The dataset contains the most frequent words from the entire battery of tests

(their average frequency is 129,088 documents in the Stanford corpus and 775 documents in TASA).

As before, GLSA was the best technique on the TASA corpus (65% correlation for both GLSA-MW and GLSA-SW) and PMI won on the larger corpus (71%). Moreover, GLSA was worse on Stanford than on TASA.

RGP (Rohde et al., 2006). Rohde et al. (2006) ran a survey to determine semantic similarity among 400 pairs of words that spanned various types of lexical relationships (e.g., morphologically related words such as *teacher – teachers*, taxonomically related words such as *apple – pear*, synonyms such as *dog – hound*, phonetically similar words *pond – ponder*). These words belonged to different word types. On average each word was rated by 33 participants. The average word frequency was 477 documents for TASA and 70,382 for Stanford.

On the RGP, the best correlation was achieved by LSA, when we used the TASA corpus. On the Stanford corpus, both GLSA methods were slightly better than PMI.

N (Nelson et al., 2005). Nelson et al. (2005) collected similarity ratings for 1016 word pairs on a scale from 1 to 7. Ninety-four people produced these ratings. Some samples include *accomplish – succeed, afraid – dark, alike – similar, absence – tardy*. The average word frequency was 92,079 for the Stanford corpus and 740 for TASA.

LSA tended to perform best again on this high-frequency word set, using the TASA corpus; however, the correlation with the human raters was very low (25%), as was the correlation achieved by the other two methods. Using a larger corpus did not improve the correlations. The size of this dataset was very large, compared with the other datasets; moreover, all the words in this dataset were rated as pretty similar by people (between 2.5 and 6.5). Moreover, the human inter-rater correlation was very low (on average 34%, maximum 44%). When we looked at how well the individual subjects were correlated with the average ratings (which were compared against the results from LSA, PMI, and GLSA given in Table 3), we obtained that the maximum subject-average correlation was 63% and the minimum 14%. On average, each individual rater correlated to the average ratings to a level of only 42%. Thus, the results of the three methods were definitely within the range of human correlations.

Discussion

Whereas GLSA was consistently better on the synonymy tests for both corpora (with GLSA-MW and GLSA-SW producing very close results), the picture is a little different on the word similarity tests. GLSA still tends to be the best when the TASA corpus is used, but PMI is usually better on the Stanford corpus. Moreover, for word similarity tests,

Test	TASA	Stanford
TOEFL	423	56656
ESL	501	71559
RD	480	24773
WS353	775	129088
MC	512	42435
R	524	40466
RG	354	29938
RGP	477	70382
N	739	92079

Table 4: Average word frequencies for each test for TASA and Stanford corpora.

Frequency	GLSA (TASA)	GLSA (Stanford)	PMI (TASA)	PMI (Stanford)	LSA
TASA	-0.53	-0.74	-0.05	-0.29	-0.42
Stanford	-0.39	-0.54	-0.06	-0.23	-0.31

Table 5: Correlations of test performance with average word frequency.

the performance of GLSA actually deteriorates on the Stanford corpus when compared to the TASA corpus.

It is useful to meditate a little on the difference between synonymy tests and word similarity tests. Synonymy looks at words which have high similarity and can be substituted to each other, whereas similarity is broader. Two related concepts can be similar, but not necessarily synonymous. It seems that, at least on a large corpus, GLSA is better suited to capture synonymy (for instance, for applications that need to expand queries by generating synonym terms), whereas PMI is more appropriate if one needs to generate terms in the same knowledge domain (e.g., for automatic highlighting of terms related to a certain topic).

An interesting observation is that the three measures correlate very well (albeit negatively) with the average frequency of the words in the test: the lower the average frequency of words in a text, the better the performance on that test. Table 4 shows how the average word frequency varies among tests; Table 5 indicates how these frequencies correlate with the performance number in Table 3. PMI correlates least with frequencies, whereas GLSA correlates most. It is possible that the more frequent words have more meanings attached to them, and thus there may be harder to capture the correct similarity pattern for those words.

With respect to the number of factors, we did not see any systematic tendency for the performance to improve around a certain magic number. The number of factors that produces optimum performance varies quite a bit for GLSA, although we seem find the optimum performance most often in the range 0 – 110 on TASA. However, for LSA the best performance is obtained usually for a number of factors in the range 250 – 300.

For GLSA, we have also experimented with a variable number of terms (up to 32,000). The performance increases with the number of terms, reaching a plateau around 16,000 terms. This plateau probably explains why the performance for the two GLSA techniques (GLSA-MW and GLSA-SW) is not different, since they use approximately the same term list.

Another manipulation that we carried out for GLSA and PMI, but that we do not report here in detail, was the use of stemming in the corpus. Consistent with previous research on LSA (Nakov, Valchanova, & Angelova, 2003), stemming did not improve substantially the performance of GLSA or PMI. (LSA uses the unstemmed TASA version.) When using the stemmed and the unstemmed versions of TASA, we mostly noticed an increase in performance for the stemmed TASA for both PMI and GLSA. For this corpus GLSA seems less sensitive to stemming than PMI though (9% average increase for PMI and only 1% average increase for GLSA), probably because by stemming PMI gets more examples in a small corpus. Most of the time GLSA's performance was lower on the stemmed Stanford corpus than on the unstemmed Stanford corpus (on average 1.5% lower); this is also true for PMI (on average 1% lower).

Related Work

Since LSA was introduced in 1997 (Landauer & Dumais, 1997), there have been several studies that have proposed alternative methods and attempted to compare them to LSA and to the 65% accuracy baseline that LSA had established on TOEFL. When Turney (2001) proposed PMI as a similarity measure, he compared its performance with that of LSA on two synonymy tests: TOEFL and ESL, in which participants choose the best synonym for a word among four possible choices. Two different corpora were used for the two measures: for PMI, the 350 million pages forming the Alta Vista database at the time; for LSA, the encyclopedia corpus available from the LSA web page. Turney used several ways of counting co-occurrence: (1) within a 10-word window, (2) in the entire document, (3) within a 10-word window, but correcting for the presence of word *not* (which may signal an antonym). With the Alta Vista corpus, PMI produced better results on TOEFL when the small window was used (approximately 73% accuracy). PMI also produced scores ranging from 62% to 66% accuracy when the 10-word window was used.

Terra and Clarke (2003) compared PMI with other statistical measures of word sim-

ilarity (χ^2 tests, likelihood ratio, average mutual information, and other techniques that took word context into account) on TOEFL. The corpus used was made of 77 million documents crawled from the web. PMI scored best, with an accuracy of 81%, using a window size of 16-32 words. Terra and Clarke (2003) also tested smaller size sub-corpora of their corpus and decided that the impact of the corpus size reaches an asymptote (i.e., no better performance is obtained by increasing the corpus), but the point where it occurs varies with each evaluation metric (e.g., for TOEFL and PMI, the asymptote was reached around 800 gigabytes out of a total corpus of 1 terabyte).

When Matveeva et al. (2005) introduced GLSA, they measured the performance of their technique on TOEFL and compared it with the precision of PMI. They ran PMI and GLSA on the same collection of documents (English Gigaword collection of New York Times articles), containing 1,1 million documents. For GLSA, the term list (vocabulary) V contained the words of TOEFL and the 2000 most frequent words from the corpus. GLSA obtained scores of 86% accuracy on the synonymy test, as opposed to PMI, which yielded about 70% accuracy on the same collection.

Kaur and Hornof (2005) compared PMI, LSA, and WordNet (Miller, 1995) on a task that targeted information scent directly: they looked at how well these measures predicted users' navigation choices. For LSA they used TASA corpus available from the LSA web page, as well as two small (containing approximately 4000 and 13,000 documents) domain-specific corpora. For PMI they used a static corpus of 77 million web pages, obtained by crawling the web. PMI performed best of all measures, achieving an accuracy of predicting people's choices of about 55%, whereas LSA scored around 40%. For LSA, Kaur and Hornof (2005) also discovered that the best performing corpus was domain-specific; however, there was no improvement in using a large domain-specific corpus versus a smaller one.

The literature thus indicates that PMI performs very well on a variety of tests and on different corpora. Although Matveeva et al. (2005) did compare GLSA and PMI on the same corpus, no previous work has compared all three techniques on one corpus. Moreover, the TASA corpus is often used in different LSA applications that use the University of Colorado interface to compute LSA values, so it is useful to see how other techniques perform on that corpus. Although the results of Terra and Clarke (2003) and Kaur and Hornof (2005) seem to indicate that a small corpus may be as good as a larger one, there was no systematic comparison between two corpora that differ in their qualities as much as TASA and the Stanford corpus do (TASA being carefully put together to reflect the college students' vocabulary, whereas Stanford is generated by a web crawl).

Conclusions

Information scent is a major construct in navigation tasks that is often estimated by measures of semantic similarity. We examined three such measures (LSA, PMI, and GLSA) and how they compare to each other on several synonymy and word similarity tests.

In terms of space and time complexity, PMI is the most lightweight technique. When compared to LSA, GLSA has the advantage of storing a relatively small (word-by-word) matrix and thus, it is less dependent computationally on the size of the corpus used, whereas the size of the corpus used by LSA is limited by the size of the word-by-document matrix that must be stored in memory. However, unlike LSA, each time the similarity of a new word (not on the term list) is needed, the entire GLSA computation needs to be carried again with an extended term list.

Since all these measures rely on word co-occurrence in a large document set, it is important to evaluate them on the same corpus. Another question of interest is whether a large, easily accessible web-based corpus, such as the Stanford corpus, can substitute a carefully formed corpus such as TASA. The current study found that GLSA outperforms LSA and PMI on the TASA corpus. On the Stanford corpus, GLSA is still best for capturing synonymy relationships between words; however, PMI is better if broader similarity judgments are needed (e.g., for finding all terms related to a certain topic). PMI works best with a large web-based corpus, possibly because it is exclusively based on word co-occurrences with no attempt to remove any noise that may be present in such data, and thus a large corpus may include more representative examples. For word similarity tests, GLSA works better with a carefully formed corpus such as TASA, but for synonymy tests, Stanford leads to better performance. In conclusion, PMI seems like a winning technique for broader judgments of word similarity: it can be easily computed from a web-based corpus and the performance is very good. GLSA (which essentially combines PMI with SVD) may be better suited on a small, representative corpus such as TASA. Thus, GLSA is best for applications that use a large corpus and need strict estimates of synonymy, or for applications that involve small, domain-specific corpora.

Acknowledgments

We would like to thank TASA and Robert Millard for giving us permission to use their corpus; Marilyn Hughes Blackmon for making the TASA corpus available to us; Doug Rohde and Doug Nelson for sharing their similarity data sets. Portions of this research have been funded by ARDA/NIMD Contract No. MDA904-03-C-0404.

References

Blackmon, M., Kitajima, M., & Polson, P. (2005). Web interactions: Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In *Proc. chi 2005.*

Budiu, R., & Pirolli, P. (2006). Navigation in degree-of-interest trees. In *Proceedings of Advance Visual Interfaces conference (avi 2006).*

Card, S., Mackinlay, J., & Schneiderman, B. (1999). *Information visualization.* Morgan Kaufmann.

Chi, E., Rosien, A., & Suppattanasiri, G. (2003). The bloodhound project: Automating the discovery of web usability issues using the infoscent simulator. In *Proc. chi 2003.*

Cho, J., Garcia-Molina, H., Haveliwala, T., Lam, W., Paepcke, A., Raghavan, S., & Wesley, G. (2004). *Stanford webbase components and applications* (Tech. Rep.). Stanford University.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Z., S., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions of Information Systems*, *20*(1), 116-131.

Jarmasz, M., & Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Proceedings of conference on recent advances in natural language processing (ranlp 2003)* (p. 212-219). Borovets, Bulgaria.

Kaur, I., & Hornof, A. J. (2005). A comparison of lsa, wordnet and pmi-ir for predicting user click behavior. In *Proc. chi 2005.* ACM Press.

Landauer, T. K., & Dumais, S. (1997). A solution to plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 211-240.

Landauer, T. K., Foltz, P., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-284.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

Matveeva, I., Levow, G., Farahat, A., & Royer, C. (2005). Terms representation with generalized latent semantic analysis. In *Proc. ranlp 2005.*

Miller, G., & Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1-28.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Commun. ACM*, *38*(11), 39–41.

Nakov, P., Valchanova, E., & Angelova, G. (2003). Towards deeper understanding of the lsa performance. In *Proc. recent advances in natural language processing* (p. 311-318). Borovetz, Bulgaria.

Nelson, D. L., Dyrdal, G. M., & Goodmon, L. B. (2005). What is preexisting strength? predicting free association probabilities, similarity ratings, and cued recall probabilities. *Psychonomic Bulletin & Review*, *12*, 711-719.

Niwa, Y., & Nitta, Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th conference on computational linguistics* (pp. 304–309). Morristown, NJ, USA: Association for Computational Linguistics.

Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive science*, *29*(3), 343-373.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*.

Pirolli, P., Card, S., & Van Der Wege, M. (2000). The effect of information scent on searching information visualizations of large tree structures. In *Proc. avi 2000.*

Resnick, P. (1995). Using information content to evaluate semantic similarity. In *Proc. ijcai 1995.*

Rohde, D., Gonnerman, L., & Plaut, D. (2006). *An improved model of semantic similarity based on lexical co-occurence.* (Manuscript submitted to *Cognitive Science*)

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627-633.

Spool, J., Perfetti, C., & Brittan, D. (2004). Designing for the scent of information. *UI Engineering*.

Terra, E., & Clarke, C. L. A. (2003). Frequency estimates for statistical word similarity measures. In *Naacl '03: Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology* (pp. 165–172). Morristown, NJ, USA: Association for Computational Linguistics.

Turney, P. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proc. emcl 2001*.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide.* Touchstone Applied Science Associates (TASA), Inc.