New Experiments in Distributional Representations of Synonymy

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow,
Sadik Kapadia, Richard Rohwer, Zhiqiang Wang
HNC Software, LLC
3661 Valley Centre Drive
San Diego, CA 92130, USA
{DayneFreitag,MatthiasBlume,JohnByrnes,EdChow,
SadikKapadia,RichardRohwer,ZhiqiangWang}@fairisaac.com

Abstract

Recent work on the problem of detecting synonymy through corpus analysis has used the Test of English as a Foreign Language (TOEFL) as a benchmark. However, this test involves as few as 80 questions, prompting questions regarding the statistical significance of reported results. We overcome this limitation by generating a TOEFL-like test using WordNet, containing thousands of questions and composed only of words occurring with sufficient corpus frequency to support sound distributional comparisons. Experiments with this test lead us to a similarity measure which significantly outperforms the best proposed to date. Analysis suggests that a strength of this measure is its relative robustness against polysemy.

1 Introduction

Many text applications are predicated on the idea that shallow lexical semantics can be acquired through corpus analysis. Harris articulated the expectation that words with similar meanings would be used in similar contexts (Harris, 1968), and recent empirical work involving large corpora has borne this out. In particular, by associating each word with a distribution over the words observed in its context, we can distinguish synonyms from non-synonyms with fair reliability. This capability may be exploited to generate corpus-based thesauri automatically (Lin, 1998), or used in any other application of text that might benefit from a measure of lexical semantic similarity. And synonymy is a logical first step in a broader research program that seeks to account for natural language semantics through distributional means.

Previous research into corpus-analytic approaches to synonymy has used the Test of English as a Foreign Language (*TOEFL*). The TOEFL consists of 300 multiple-choice question, each question involving five words: the problem or target word and four response words, one of which is a synonym of the target. The objective is to identify the synonym (call this the *answer* word, and call the other response words *decoys*). In the context of research into lexical semantics, we seek a distance function which as reliably as possible orders the answer word in front of the decoys.

Landauer and Dumais first proposed the TOEFL as a test of lexical semantic similarity and reported a score of 64.4% on an 80-question version of the TOEFL, a score nearly identical to the average score of human test takers (Landauer and Dumais, 1997). Subsequently, Sahlgren reported a score of 72.0% on the same test using “random indexing” and a different training corpus (Sahlgren, 2001). By analyzing a much larger corpus, Ehlert was able to score 82% on a 300-question version of the TOEFL, using a simple distribution over contextual words (Ehlert, 2003).

While success on the TOEFL does not immediately guarantee success in real-word applications requiring lexical similarity judgments, the scores have an intuitive appeal. They are easily interpretable, and the expected performance of a random

Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL),
pages 25–32, Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

guesser (25%) and typical human performance are both known. Nevertheless, the TOEFL is problematic in at least two ways. On the one hand, because it involves so few questions, conclusions based on the TOEFL regarding closely competing approaches are suspect. Even on the 300-question TOEFL, a score of 82% is accurate only to within plus or minus 4% at the 95% confidence level. The other shortcoming is a potential mis-match between the test vocabulary and the corpus vocabulary. Typically, a substantial number of questions include words observed too infrequently in the training corpus for a semantic judgment to be made with any confidence.

We seek to overcome these difficulties by generating TOEFL-like tests automatically from WordNet (Fellbaum, 1998). While WordNet has been used before to evaluate corpus-analytic approaches to lexical similarity (Lin, 1998), the metric proposed in that study, while useful for comparative purposes, lacks an intuitive interpretation. In contrast, we emulate the TOEFL using WordNet and inherit the TOEFL's easy interpretability.

Given a corpus, we first derive a list of words occurring with sufficient marginal frequency to support a distributional comparison. We then use WordNet to generate a large set of questions identical in format to those in the TOEFL. For a vocabulary of reasonable size, this yields questions numbering in the thousands. While the resulting questions differ in some interesting ways from those in the TOEFL (see below), their sheer number supports more confident conclusions. Beyond this, we can partition them by part of speech or degree of polysemy, enabling some analyses not supported by the original TOEFL.

2 The Test

To generate a TOEFL-like test from WordNet, we perform the following procedure once each for nouns, verbs, adjectives and adverbs. Given a list of candidate words, we produce one test question for every ordered pair of words appearing together in any synset in the respective WordNet part-of-speech database. Decoy words are chosen at random from among other words in the database that do not have a synonymy relation with either word in the pair. For convenience, we will call the resulting test the *WordNet-based synonymy test* (WBST).

technology:	
A. engineering	B. difference
C. department	D. west
stadium:	
A. miss	B. hockey
C. wife	D. bowl
string:	
A. giant	B. ballet
C. chain	D. hat
trial:	
A. run	B. one-third
C. drove	D. form

Table 1: Four questions chosen at random from the noun test. Answers are A, D, C, and A.

We take a few additional steps in order to increase the resemblance between the WBST and the TOEFL. First, we remove from consideration any stop words or inflected forms. Note that whether a particular wordform is inflected is a function of its presumed part of speech. The word "indicted" is either an inflected verb (so would not be used as a word in a question involving verbs) or an uninflected adjective. Second, we rule out pairs of words that are too similar under the string edit distance. Morphological variants often share a synset in WordNet. For example, "group" and "grouping" share a nominal sense. Questions using such pairs appear trivial to human test takers and allow stemming shortcuts.

In the experiments reported in this paper, we used WordNet 1.7.1. Our experimental corpus is the North American News corpus, which is also used by Ehlert (2003). We include as a candidate test word any word occurring at least 1000 times in the corpus (about 15,000 words when restricted to those appearing in WordNet). Table 1 shows four sample questions generated from this list out of the noun database. In total, this procedure yields 9887 noun, 7398 verb, 5824 adjective, and 461 adverb questions, a total of 23,570 questions.[1]

This procedure yields questions that differ in some interesting ways from those in the TOEFL. Most notable is a bias in favor of polysemous terms. The number of times a word appears as either the target or the answer word is proportional to the number of synonyms it has in the candidate list. In contrast,

[1] This test is available as `http://www.cs.cmu.edu/~dayne/wbst-nanews.tar.gz`.

decoy words are chosen at random, so are less polysemous on average.

3 The Space of Solutions

Given that we have a large number of test questions composed of words with high corpus frequencies, we now seek to optimize performance on the WBST. The solutions we consider all start with a word-conditional context frequency vector, usually normalized to form a probability distribution. We answer a question by comparing the target term vector and each of the response term vectors, choosing the "closest."

This problem definition excludes a common class of solutions to this problem, in which the closeness of a pair of terms is a statistic of the co-occurrence patterns of the specific terms in question. It has been shown that measures based on the pointwise mutual information (PMI) between question words yield good results on the TOEFL (Turney, 2001; Terra and Clarke, 2003). However, Ehlert (2003) shows convincingly that, for a fixed amount of data, the distributional model performs better than what we might call the pointwise co-occurrence model. Terra and Clark (2003) report a top score of 81.3% on an 80-word version of the TOEFL, which compares favorably with Ehlert's best of 82% on a 300-word version, but their corpus is approximately 200 times as large as Ehlert's.

Note that these two approaches are complementary and can be combined in a supervised setting, along with static resources, to yield truly strong performance (97.5%) on the TOEFL (Turney et al., 2003). While impressive, this work begs an important question: Where do we obtain the training data when moving to a less commonly taught language, to say nothing of the comprehensive thesauri and Web resources? In this paper, we focus on shallow methods that use only the text corpus. We are interested less in optimizing performance on the TOEFL than in investigating the validity and limits of the distributional hypothesis, and in illuminating the barriers to automated human-level lexical similarity judgments.

3.1 Definitions of Context

As in previous work, we form our context distributions by recording word-conditional counts of feature occurrences within some fixed window of a reference token. In this study, features are just unnormalized tokens, possibly augmented with direction and distance information. In other words, we do not investigate the utility of stemming. Similarly, except where noted, we do not remove stop words.

All context definitions involve a window size, which specifies the number of tokens to consider on *either* side of an occurrence of a reference term. It is always symmetric. Thus, a window size of one indicates that only the immediately adjacent tokens on either side should be considered. By default, we bracket a token sequence with pseudo-tokens "`<bos>`" and "`<eos>`".[2]

Contextual tokens in the window may be either observed or disregarded, and the policy governing which to admit is one of the dimensions we explore here. The decision whether or not to observe a particular contextual token is made before counting commences, and is not sensitive to the circumstances of a particular occurrence (e.g., its participation in some syntactic relation (Lin, 1997; Lee, 1999)). When a contextual token is observed, it is always counted as a single occurrence. Thus, in contrast with earlier approaches (Sahlgren, 2001; Ehlert, 2003), we do not use a weighting scheme that is a function of distance from the reference token.

Once we have chosen to observe a contextual token, additional parameters govern whether counting should be sensitive to the side of the reference token on which it occurs and how distant from the reference token it is. If the *strict direction* parameter is true, a left occurrence is distinguished from a right occurrence. If *strict distance* is true, occurrences at distinct removes (in number of tokens) are recorded as distinct event types.

3.2 Distance Measures

The product of a particular context policy is a co-occurrence matrix N, where the contents of a cell $N_{w,c}$ is the number of times context c is observed to occur with word w. A row of this matrix (N_w) is

[2] In this paper, a sequence is a North American News segment delimited by the `<p>` tag. Nominally paragraphs, most of these segments are single sentences.

therefore a word-conditional context frequency vector. In comparing two of these vectors, we typically normalize counts so that all cells in a row sum to one, yielding a word-conditional distribution over contexts $P(c|w)$ (but see the Cosine measure below).

We investigate some of the distance measures commonly employed in comparing term vectors. These include:

Manhattan	$\sum_i \lvert P(c_i \lvert w_1) - P(c_i \lvert w_2) \rvert$
Euclidean	$\sqrt{\sum_i [P(c_i \lvert w_1) - P(c_i \lvert w_2)]^2}$
Cosine	$\frac{\sum_i N_{w_1,c_i} N_{w_2,c_i}}{\lvert\lvert N_{w_1} \rvert\rvert \cdot \lvert\lvert N_{w_2} \rvert\rvert}$

Note that whereas we use probabilities in calculating the Manhattan and Euclidean distances, in order to avoid magnitude effects, the Cosine, which defines a different kind of normalization, is applied to raw number counts.

We also avail ourselves of measures suggested by probability theory. For $\delta \in (0, 1)$ and word-conditional context distributions p and q, we have the so-called δ-divergences (Zhu and Rohwer, 1998):

$$D_\delta(p, q) := \frac{1 - \sum p^\delta q^{1-\delta}}{\delta(1 - \delta)} \tag{1}$$

Divergences D_0 and D_1 are defined as limits as $\delta \to 0$ and $\delta \to 1$:

$$D_1(p, q) = D_0(q, p) = \sum p \log \frac{p}{q}$$

In other words, $D_1(p, q)$ is the KL-divergence of p from q. Members of this divergence family are in some sense preferred by theory to alternative measures. It can be shown that the δ-divergences (or divergences defined by combinations of them, such as the Jensen-Shannon or “skew” divergences (Lee, 1999)) are the only ones that are robust to redundant contexts (i.e., only divergences in this family are *invariant*) (Csiszár, 1975).

Several notions of lexical similarity have been based on the KL-divergence. Note that if any $q_i = 0$, then $D_1(p, q)$ is infinite; in general, the KL-divergence is very sensitive to small probabilities, and careful attention must be paid to smoothing if it is to be used with text co-occurrence data. The Jensen-Shannon divergence—an average of the divergences of p and q from their mean distribution—does not share this sensitivity and has previously been used in tests of lexical similarity (Lee, 1999). Furthermore, unlike the KL-divergence, it is symmetric, presumably a desirable property in this setting, since synonymy is a symmetric relation, and our test design exploits this symmetry.

However, $D_{1/2}(p, q)$, the Hellinger distance[3], is also symmetric and robust to small or zero estimates. To our knowledge, the Hellinger distance has not previously been assessed as a measure of lexical similarity. We experimented with both the Hellinger distance and Jensen-Shannon (JS) divergence, and obtained close scores across a wide range of parameter settings, with the Hellinger yielding a slightly better top score. We report results only for the Hellinger distance below. As will be seen, neither the Hellinger nor the JS divergence are optimal for this task.

In pursuit of synonymy, Ehlert (2003) derives a formula for the probability of the target word given a response word:

$$\begin{aligned} P(w_1|w_2) &= \frac{\sum_i P(w_1|c_i) P(w_2|c_i) P(c_i)}{P(w_2)} && (2) \\ &= P(w_1) \sum_i \frac{P(c_i|w_1) P(c_i|w_2)}{P(c_i)} && (3) \end{aligned}$$

The second line, which fits more conveniently into our framework, follows from the first (Ehlert's expression) through an application of Bayes Theorem. While this measure falls outside the class of δ-divergences, our experiments confirm its relative strength on synonymy tests.

It is possible to unify the δ-divergences with Ehlert's expression by defining a broader class of measures:

$$D_{\delta,\gamma,\alpha}(p, q) = 1 - \sum_i c_i^{-\alpha} p_i^\delta q_i^\gamma \tag{4}$$

where c_i is the marginal probability of a single context, and p_i and q_i are its respective word-conditional probabilities. Since, in the context of a given question, $P(w_1)$ does not change, maximizing the expression in Equation 3 is the same as minimizing $D_{1,1,1}$. $D_{\delta,(1-\delta),0}$ recovers the δ divergences up to a constant multiple, and $D_{1,1,0}$ provides the complement of the familiar inner-product measure.

[3] Actually, $D_{1/2}(p, q)$ is four times the square of the Hellinger distance.

4 Evaluation

We experimented with various distance measures and context policies using the full North American News corpus. We count approximately one billion words in this corpus, which is roughly four times the size of the largest corpus considered by Ehlert.

Except where noted, the numbers reported here are the result of taking the full WBST, a total of 23,570 test questions. Given this number of questions, scores where most of the results fall are accurate to within plus or minus 0.6% at the 95% confidence level.

4.1 Performance Bounds

In order to provide a point of comparison, the paper's authors each answered the same random sample of 100 questions from each part of speech. Average performance over this sample was 88.4%. The one non-native speaker scored 80.3%. As will be seen, this is better than the best automated result.

The expected score, in the absence of any semantic information, is 25%. However, as noted, target and answer words are more polysemous than decoy words on average, and this can be exploited to establish a higher baseline. Since the frequency of a word is correlated with its polysemy, a strategy which always selects the most frequent word among the response words yields 39.2%, 34.5%, 29.1%, and 38.0% on nouns, verbs, adjectives, and adverbs, respectively, for an average score of 35.2%.

4.2 An Initial Comparison

Table 2 displays a basic comparison of the distance measures and context definitions enumerated so far. For each distance measure (Manhattan, Euclidean, Cosine, Hellinger, and Ehlert), results are shown for window sizes 1 to 4 (columns). Results are further sub-divided according to whether strict direction and distance are false (*None*), only strict direction is true (*Dir*), or both strict direction and strict distance are true (*Dir+Dist*). In bold is the best score, along with any scores indistinguishable from it at the 95% confidence level.

Notable in Table 2 are the somewhat depressed scores, compared with those reported for the TOEFL. Ehlert reports a best score on the TOEFL of 82%, whereas the best we are able to achieve on the WBST is 67.6%. Although there are differences in some of the experimental details (Ehlert employs a triangular window weighting and experiments with stemming), these probably do not account for the discrepancy. Rather, this appears to be a harder test than the TOEFL—despite the fact that all words involved are seen with high frequency.

		Window Size			
		1	2	3	4
Manh	None	54.2	58.8	60.4	60.6
	Dir	54.3	58.5	60.3	60.8
	Dir+Dist	–	57.3	58.8	58.9
Euc	None	42.9	45.3	46.6	47.6
	Dir	43.2	45.7	46.8	47.6
	Dir+Dist	–	44.9	45.3	45.6
Cos	None	44.9	46.7	47.6	48.3
	Dir	46.2	48.0	48.6	49.2
	Dir+Dist	–	48.0	48.4	48.5
Hell	None	57.9	62.3	62.2	61.0
	Dir	57.2	62.6	63.3	61.8
	Dir+Dist	–	61.2	61.7	61.1
Ehl	None	64.0	66.2	66.2	65.7
	Dir	63.9	66.9	**67.6**	**67.1**
	Dir+Dist	–	66.4	**67.2**	**67.5**

Table 2: Accuracy on the WBST: an initial comparison of distance measures and context definitions.

It is hard to escape the conclusion that, in pursuit of high scores, choice of distance measure is more critical than the specific definition of context. All scores returned by the Ehlert metric are significantly higher than any returned by other distance measures. Among the Ehlert scores, there is surprising lack of sensitivity to context policy, given a window of size 2 or larger.

Although the Hellinger distance yields scores only in the middle of the pack, it might be that other divergences from the δ-divergence family, such as the KL-divergence, would yield better scores. We experimented with various settings of δ in Equation 1. In all cases, we observed bell-shaped curves with peaks approximately at $\delta = 0.5$ and locally worst performance with values at or near 0 or 1. This held true when we used maximum likelihood estimates, or under a simple smoothing regime in which all cells of the co-occurrence matrix were initialized with various fixed values. It is possible that numerical issues are nevertheless partly responsible for the poor showing of the KL-divergence. However, given the symmetry of the synonymy relation, it would be surprising if some value of δ far from 0.5 was ultimately shown to be best.

4.3 The Importance of Weighting

The Ehlert measure and the cosine are closely related—both involve an inner product between vectors—yet they return very different scores in Table 2. There are two differences between these methods, normalization and vector element weighting. We presume that normalization does not account for the large score difference, and attribute the discrepancy, and the general strength of the Ehlert measure, to importance weighting.

In information retrieval, it is common to take the cosine between vectors where vector elements are not raw frequency counts, but counts weighted using some version of the "inverse document frequency" (IDF). We ran the cosine experiment again, this time weighting the count of context i by $log(D/d_i)$, where D is the number of rows in the count matrix N and d_i is the number of rows containing a non-zero count for context i. The results confirmed our expectation. The performance of "CosineIDF" for a window size of 3 with strict direction was 64.0%, which is better than Hellinger but worse than the Ehlert measure. This was the best result returned for "CosineIDF."

4.4 Optimizing Distance Measures

Both the Hellinger distance and the Ehlert measure are members of the family of measures defined by Equation 4. Although there are theoretical reasons to prefer each to neighboring members of the same family (see the discussion following Equation 1), we undertook to validate this preference empirically. We conducted parameter sweeps of α, δ, and γ, first exploring members of the family $\delta = \gamma$, of which both Hellinger and Ehlert are members. Specifically, we explored the space between $\delta = \gamma = 0.5$ and $\delta = \gamma = 1$, first in increments of 0.1, then in increments of 0.01 around the approximate maximum, in all cases varying α widely.

This experiment clearly favored a region midway between the Hellinger and Ehlert measures. We identified $\delta = \gamma = 0.75$, with $\alpha = 1.1$ as the approximate midpoint of this optimal region. We next varied δ and γ independently around this point. This resulted in no improvement to the score, confirming our expectation that some point along $\delta = \gamma$ would be best. For the sake of brevity, we will refer to this best point ($D_{0.75,0.75,1.1}$) as the "Optimal" measure. As Table 3 indicates, this measure is significantly better than the Ehlert measure, or any other measure investigated here.

	Noun	Verb	Adj	Adv	All
Ehlert	71.6	57.2	73.4	72.5	67.6
Optimal	75.8	63.8	76.4	76.6	72.2

Table 3: Comparison between the Ehlert measure and the "optimal" point in the space of measures defined by Equation 4 ($\delta = \gamma = 0.75$, $\alpha = 1.1$), by part of speech. Context policy is window size 3 with strict direction.

This clear separation between Ehlert and Optimal does not hold for the original TOEFL. Using the same context policy, we applied these measures to 298 of the 300 questions used by Ehlert (all questions except those involving multi-word terms, which our framework does not currently support). Optimal returns 84.2%, while Ehlert's measure returns 83.6%, which is slightly better than the 82% reported by Ehlert. The two results are not distinguishable with any statistical significance.

Interesting in Table 3 is the range of scores seen across parts of speech. The variation is even wider under other measures, the usual ordering among parts of speech being (from highest to lowest) adverb, adjective, noun, verb. In Section 4.6, we attempt to shed some light on both this ordering and the close outcome we observe on the TOEFL.

4.5 Optimizing Context Policy

It is certain that not every contextual token seen within the co-occurrence window is equally important to the detection of synonymy, and probable that some such tokens are useless or even detrimental. On the one hand, the many low-frequency events in the tails of the context distributions consume a lot of space, perhaps without contributing much information. On the other, very-high-frequency terms are typically closed-class and stop words, possibly too common to be useful in making semantic distinctions. We investigated excluding words at both ends of the frequency spectrum.

We experimented with two kinds of exclusion policies: one excluding the k most frequent terms, for k ranging between 10 and 200; and one excluding terms occurring fewer than k times, for k ranging from 3 up to 100. Both Ehlert and Optimal were largely invariant across all settings; no statistically significant improvements or degradations were observed. Optimal returned scores ranging from 72.0%, when contexts with marginal frequency fewer than 100 were ignored, up to 72.6%, when the 200 most frequent terms were excluded.

Note there is a large qualitative difference between the two exclusion procedures. Whereas we exclude only at most 200 words in the high-frequency experiment, the number of terms excluded in the low-frequency experiment ranges from 939,496 (less than minimum frequency 3) to 1,534,427 (minimum frequency 100), out of a vocabulary containing about 1.6 million terms. Thus, it is possible to reduce the expense of corpus analysis substantially without sacrificing semantic fidelity.

4.6 Polysemy

We hypothesized that the variation in scores across part of speech has to do with the average number of senses seen in a test set. Common verbs, for example, tend to be much more polysemous (and syntactically ambiguous) than common adverbs. WordNet allows us to test this hypothesis.

We define the *polysemy level* of a question as the sum of the number of senses in WordNet of its target and answer words. Polysemy levels in our question set range from 2 up to 116. Calculating the average polysemy level for questions in the various parts of speech—5.1, 6.7, 7.5, and 10.4, for adverbs, adjectives, nouns, and verbs, respectively—provides support for our hypothesis, inasmuch as this ordering aligns with test scores. By contrast, the average polysemy level in the TOEFL, which spans all four parts of speech, is 4.6.

Plotting performance against polysemy level helps explain why Ehlert and Optimal return roughly equivalent performance on the original TOEFL. Figure 1 plots the Ehlert and Optimal measures as a function of the polysemy level of the questions. To produce this plot, we grouped questions according to polysemy level, creating many smaller tests, and scored each measure on each test separately.

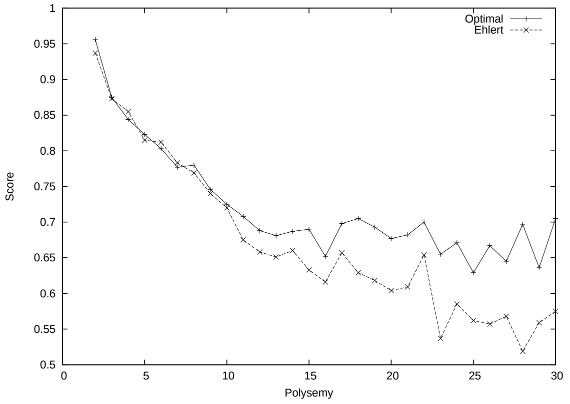

Figure 1: Score as a function of polysemy level.

At low polysemy levels, the Ehlert and Optimal measures perform equally well. The advantage of Optimal over Ehlert appears to lie specifically in its relative strength in handling polysemous terms.

5 Discussion

Specific conclusions regarding the “Optimal” measure are problematic. We do not know whether or to what extent this particular parameter setting is universally best, best only for English, best for newswire English, or best only for the specific test we have devised. We have restricted our attention to a relatively small space of similarity measures, excluding many previously proposed measures of lexical affinity (but see Weeds, et al (2004), and Lee (1999) for some empirical comparisons). Lee observed that measures from the space of invariant divergences (particularly the JS and skew divergences) perform at least as well as any of a wide variety of alternatives. As noted, we experimented with the JS divergence and observed accuracies that tracked those of the Hellinger closely. This provides a point of comparison with the measures investigated by Lee, and recommends both Ehlert’s measure and what we have called “Optimal” as credible, perhaps superior alternatives. More generally, our results argue for some form of feature importance weighting.

Empirically, the strength of Optimal on the WBST is a feature of its robustness in the presence of polysemy. Both Ehlert and Optimal are expressed as a sum of ratios, in which the numerator is a product of some function of conditional context probabilities, and the denominator is some function of the marginal probability. The Optimal exponents on both the numerator and denominator have the effect of advantaging lower-probability events, relative to Ehlert. In our test, WordNet senses are sampled uniformly at random. Perhaps its emphasis on lower probability events allows Optimal to sacrifice some fidelity on high-frequency senses in exchange for increased sensitivity to low-frequency ones.

It is clear, however, that polysemy is a critical hurdle confronting distributional approaches to lexical semantics. Figure 1 shows that, in the absence of polysemy, distributional comparisons detect synonymy quite well. Much of the human advantage over machines on this task may be attributed to an awareness of polysemy. In order to achieve performance comparable to that of humans, therefore, it is probably not enough to optimize context policies or to rely on larger collections of text. Instead, we require strategies for detecting and resolving latent word senses.

Pantel and Lin (2002) propose one such method, evaluated by finding the degree of overlap between sense clusters and synsets in WordNet. The above considerations suggest that a possibly more pertinent test of such approaches is to evaluate their utility in the detection of semantic similarity between specific polysemous terms. We expect to undertake such an evaluation in future work.

Acknowledgments. This material is based on work funded in whole or in part by the U.S. Government. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the U.S. Government.

References

I. Csiszár. 1975. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158.

B. Ehlert. 2003. Making accurate lexical semantic similarity judgments using word-context co-occurrence statistics. Master's thesis, University of California, San Diego.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Z. Harris. 1968. *Mathematical Structures of Language*. Interscience Publishers, New York.

T.K. Landauer 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(2):211–240.

L. Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th ACL*.

D. Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACL-97*, Madrid, Spain.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL98*, Montreal, Canada.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *Proceedings of KDD-02*, Edmonton, Canada.

M. Sahlgren. 2001. Vector-based semantic analysis: representing word meanings based on random labels. In *Semantic Knowledge Acquisition and Categorisation Workshop, ESSLLI 2001*, Helsinki, Finland.

E. Terra and C.L.A. Clarke. 2003. Frequency estimates for statistical word similarity measures. In *Proceedings of HLT/NAACL 2003*, Edmonton, Canada.

P.D. Turney, M.L. Littman, J. Bigham, and V. Schnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

P.D. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML-01)*.

J. Weeds, D. Weir, and D. McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of CoLing 2004*, Geneva, Switzerland.

H. Zhu and R. Rohwer. 1998. Information geometry, Bayesian inference, ideal estimates, and error decomposition. Technical Report 98-06-045, Santa Fe Institute.