CONTEXTUAL WORD SIMILARITY AND ESTIMATION FROM SPARSE DATA

Ido Dagan
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974
dagan@research.att.com

Shaul Marcus
Computer Science Department
Technion
Haifa 32000, Israel
shaul@cs.technion.ac.il

Shaul Markovitch
Computer Science Department
Technion
Haifa 32000, Israel
shaulm@cs.technion.ac.il

Abstract

In recent years there is much interest in word cooccurrence relations, such as n-grams, verb-object combinations, or cooccurrence within a limited context. This paper discusses how to estimate the probability of cooccurrences that do not occur in the training data. We present a method that makes local analogies between each specific unobserved cooccurrence and other cooccurrences that contain similar words, as determined by an appropriate word similarity metric. Our evaluation suggests that this method performs better than existing smoothing methods, and may provide an alternative to class based models.

1 Introduction

Statistical data on word cooccurrence relations play a major role in many corpus based approaches for natural language processing. Different types of cooccurrence relations are in use, such as cooccurrence within a consecutive sequence of words (n-grams), within syntactic relations (verb-object, adjective-noun, etc.) or the cooccurrence of two words within a limited distance in the context. Statistical data about these various cooccurrence relations is employed for a variety of applications, such as speech recognition (Jelinek, 1990), language generation (Smadja and McKeown, 1990), lexicography (Church and Hanks, 1990), machine translation (Brown et al., ; Sadler, 1989), information retrieval (Maarek and Smadja, 1989) and various disambiguation tasks (Dagan et al., 1991; Hindle and Rooth, 1991; Grishman et al., 1986; Dagan and Itai, 1990).

A major problem for the above applications is how to estimate the probability of cooccurrences that were not observed in the training corpus. Due to data sparseness in unrestricted language, the aggregate probability of such cooccurrences is large and can easily get to 25% or more, even for a very large training corpus (Church and Mercer, 1992). Since applications often have to compare alternative hypothesized cooccurrences, it is important to distinguish between those unobserved cooccurrences that are likely to occur in a new piece of text and those that are not. These distinctions ought to be made using the data that do occur in the corpus. Thus, beyond its own practical importance, the sparse data problem provides an informative touchstone for theories on generalization and analogy in linguistic data.

The literature suggests two major approaches for solving the sparse data problem: smoothing and class based methods. Smoothing methods estimate the probability of unobserved cooccurrences using frequency information (Good, 1953; Katz, 1987; Jelinek and Mercer, 1985; Church and Gale, 1991). Church and Gale (Church and Gale, 1991) show, that for unobserved bigrams, the estimates of several smoothing methods closely agree with the probability that is expected using the frequencies of the two words and assuming that their occurrence is independent ((Church and Gale, 1991), figure 5). Furthermore, using held out data they show that this is the probability that should be estimated by a smoothing method that takes into account the frequencies of the individual words. Relying on this result, we will use *frequency based estimation* (using word frequencies) as representative for smoothing estimates of unobserved cooccurrences, for comparison purposes. As will be shown later, the problem with smoothing estimates is that they ignore the expected degree of association between the specific words of the cooccurrence. For example, we would not like to estimate the same probability for two cooccurrences like 'eat bread' and 'eat cars', despite the fact that both 'bread' and 'cars' may have the same frequency.

Class based models (Brown et al., ; Pereira et al., 1993; Hirschman, 1986; Resnik, 1992) distinguish between unobserved cooccurrences using classes of "similar" words. The probability of a specific cooccurrence is determined using generalized parameters about the probability of class cooccur-

rence. This approach, which follows long traditions in semantic classification, is very appealing, as it attempts to capture "typical" properties of classes of words. However, it is not clear at all that unrestricted language is indeed structured the way it is assumed by class based models. In particular, it is not clear that word cooccurrence patterns can be structured and generalized to class cooccurrence parameters without losing too much information.

This paper suggests an alternative approach which assumes that class based generalizations should be avoided, and therefore eliminates the intermediate level of word classes. Like some of the class based models, we use a similarity metric to measure the similarity between cooccurrence patterns of words. But then, rather than using this metric to construct a set of word classes, we use it to identify the most specific analogies that can be drawn for each specific estimation. Thus, to estimate the probability of an unobserved cooccurrence of words, we use data about other cooccurrences that were observed in the corpus, and contain words that are similar to the given ones. For example, to estimate the probability of the unobserved cooccurrence 'negative results', we use cooccurrences such as 'positive results' and 'negative numbers', that do occur in our corpus.

The analogies we make are based on the assumption that similar word cooccurrences have similar values of mutual information. Accordingly, our similarity metric was developed to capture similarities between vectors of mutual information values. In addition, we use an efficient search heuristic to identify the most similar words for a given word, thus making the method computationally affordable. Figure 1 illustrates a portion of the similarity network induced by the similarity metric (only some of the edges, with relatively high values, are shown). This network may be found useful for other purposes, independently of the estimation method.

The estimation method was implemented using the relation of cooccurrence of two words within a limited distance in a sentence. The proposed method, however, is general and is applicable for any type of lexical cooccurrence. The method was evaluated in two experiments. In the first one we achieved a complete scenario of the use of the estimation method, by implementing a variant of the disambiguation method in (Dagan et al., 1991), for sense selection in machine translation. The estimation method was then successfully used to increase the coverage of the disambiguation method by 15%, with an increase of the overall precision compared to a naive, frequency based, method. In the second experiment we evaluated the estimation method on a data recovery task. The task simulates a typical scenario in disambiguation, and also relates to theoretical questions about redundancy and idiosyncrasy in cooccurrence data. In this evaluation, which involved 300 examples, the performance of the estimation method was by 27% better than frequency based estimation.

2 Definitions

We use the term *cooccurrence pair*, written as (x, y), to denote a cooccurrence of two words in a sentence within a distance of no more than d words. When computing the distance d, we ignore function words such as prepositions and determiners. In the experiments reported here $d = 3$.

A cooccurrence pair can be viewed as a generalization of a bigram, where a bigram is a cooccurrence pair with $d = 1$ (without ignoring function words). As with bigrams, a cooccurrence pair is directional, i.e. $(x, y) \neq (y, x)$. This captures some information about the asymmetry in the linear order of linguistic relations, such as the fact that verbs tend to precede their objects and follow their subjects.

The mutual information of a cooccurrence pair, which measures the degree of association between the two words (Church and Hanks, 1990), is defined as (Fano, 1961):

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} = \log_2 \frac{P(x|y)}{P(x)} = \log_2 \frac{P(y|x)}{P(y)} \quad (1)$$

where $P(x)$ and $P(y)$ are the probabilities of the events x and y (occurrences of words, in our case) and $P(x, y)$ is the probability of the joint event (a cooccurrence pair).

We estimate mutual information values using the Maximum Likelihood Estimator (MLE):

$$\hat{I}(x,y) = \log_2 \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)} = \log_2 \left(\frac{N}{d} \frac{f(x,y)}{f(x)f(y)}\right) \quad (2)$$

where f denotes the frequency of an event and N is the length of the corpus. While better estimates for small probabilities are available (Good, 1953; Church and Gale, 1991), MLE is the simplest to implement and was adequate for the purpose of this study. Due to the unreliability of measuring negative mutual information values in corpora that are not extremely large, we have considered in this work any negative value to be 0. We also set $\hat{I}(x, y)$ to 0 if $f(x, y) = 0$. Thus, we assume in both cases that the association between the two words is as expected by chance.

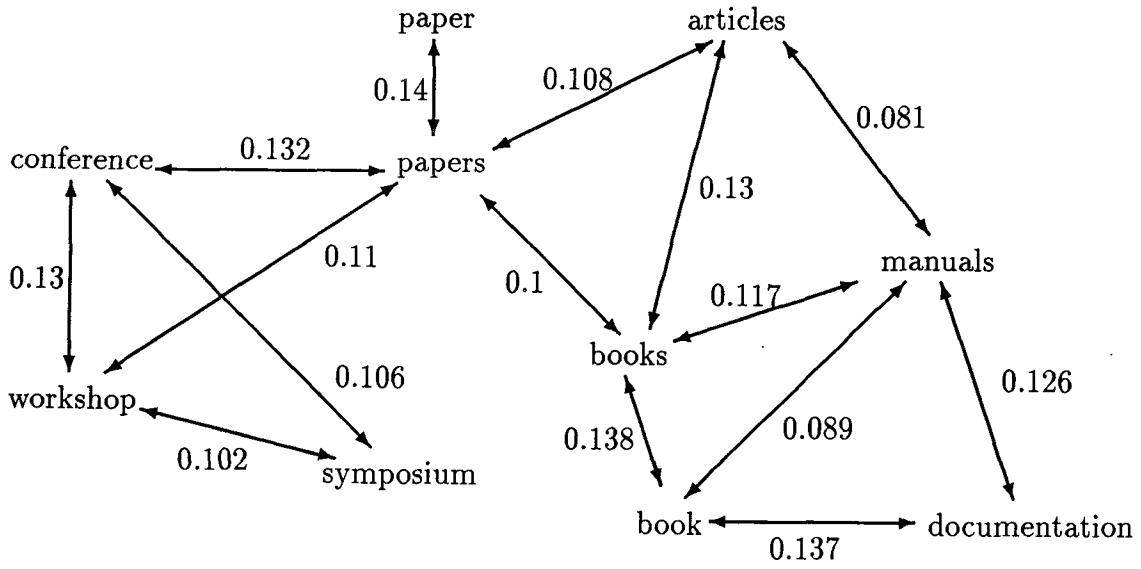

Figure 1: A portion of the similarity network.

3 Estimation for an Unobserved Cooccurrence

Assume that we have at our disposal a method for determining similarity between cooccurrence patterns of two words (as described in the next section). We say that two cooccurrence pairs, (w_1, w_2) and (w_1', w_2'), are *similar* if w_1' is similar to w_1 and w_2' is similar to w_2. A special (and stronger) case of similarity is when the two pairs differ only in one of their words (e.g. (w_1, w_2') and (w_1, w_2)). This special case is less susceptible to noise than unrestricted similarity, as we replace only one of the words in the pair. In our experiments, which involved rather noisy data, we have used only this restricted type of similarity. The mathematical formulations, though, are presented in terms of the general case.

The question that arises now is what analogies can be drawn between two similar cooccurrence pairs, (w_1, w_2) and (w_1', w_2'). Their probabilities cannot be expected to be similar, since the probabilities of the words in each pair can be different. However, since we assume that w_1 and w_1' have similar cooccurrence patterns, and so do w_2 and w_2', it is reasonable to assume that the mutual information of the two pairs will be similar (recall that mutual information measures the degree of association between the words of the pair).

Consider for example the pair *(chapter, describes)*, which does not occur in our corpus[1]. This pair was found to be similar to the pairs *(introduction, describes)*, *(book, describes)* and *(section, describes)*, that do occur in the corpus. Since these pairs occur in the corpus, we estimate their mutual information values using equation 2, as shown in Table 1. We then take the average of these mutual information values as the *similarity based estimate* for $I(chapter, describes)$, denoted as $\check{I}(chapter, describes)$[2]. This represents the assumption that the word 'describes' is associated with the word 'chapter' to a similar extent as it is associated with the words 'introduction', 'book' and 'section'. Table 2 demonstrates how the analogy is carried out also for a pair of unassociated words, such as $(chapter, knows)$.

In our current implementation, we compute $\check{I}(w_1, w_2)$ using up to 6 most similar words to each of w_1 and w_2, and averaging the mutual information values of similar pairs that occur in the corpus (6 is a parameter, tuned for our corpus. In some cases the similarity method identifies less than 6 similar words).

Having an estimate for the mutual information of a pair, we can estimate its expected frequency in a corpus of the given size using a variation of equation 2:

$$\check{f}(w_1, w_2) = \frac{d}{N} f(w_1) f(w_2) 2^{\check{I}(w_1, w_2)} \qquad (3)$$

In our example, $f(chapter) = 395$, $N = 8,871,126$ and $d = 3$, getting a similarity based estimate of $\check{f}(chapter, describes) = 3.15$. This value is much

[1] We used a corpus of about 9 million words of texts in the computer domain, taken from articles posted to the USENET news system.

[2] We use $\check{I}$ for similarity based estimates, and reserve $\hat{I}$ for the traditional maximum likelihood estimate. The similarity based estimate will be used for cooccurrence pairs that do not occur in the corpus.

(w_1, w_2)	$\check{I}(w_1, w_2)$	$f(w_1, w_2)$	$f(w_1)$	$f(w_2)$
(introduction, describes)	6.85	5	464	277
(book, describes)	6.27	13	1800	277
(section, describes)	6.12	6	923	277
Average:	6.41			

Table 1: The similarity based estimate as an average on similar pairs: $\check{I}(chapter, describes) = 6.41$

(w_1, w_2)	$\check{I}(w_1, w_2)$	$f(w_1, w_2)$	$f(w_1)$	$f(w_2)$
(introduction, knows)	0	0	464	928
(book, knows)	0	0	1800	928
(section, knows)	0	0	923	928
Average:	0			

Table 2: The similarity based estimate for a pair of unassociated words: $\check{I}(chapter, knows) = 0$

higher than the frequency based estimate (0.037), reflecting the plausibility of the specific combination of words[3]. On the other hand, the similarity based estimate for $\check{f}(chapter, knows)$ is 0.124, which is identical to the frequency based estimate, reflecting the fact that there is no expected association between the two words (notice that the frequency based estimate is higher for the second pair, due to the higher frequency of 'knows').

4 The Similarity Metric

Assume that we need to determine the degree of similarity between two words, w_1 and w_2. Recall that if we decide that the two words are similar, then we may infer that they have similar mutual information with some other word, w. This inference would be reasonable if we find that on average w_1 and w_2 indeed have similar mutual information values with other words in the lexicon. The similarity metric therefore measures the degree of similarity between these mutual information values.

We first define the similarity between the mutual information values of w_1 and w_2 relative to a single other word, w. Since cooccurrence pairs are directional, we get two measures, defined by the position of w in the pair. The *left context similarity* of w_1 and w_2 relative to w, termed $sim_L(w_1, w_2, w)$, is defined as the ratio between the two mutual information values, having the larger value in the denominator:

$$sim_L(w_1, w_2, w) = \frac{\min(I(w, w_1), I(w, w_2))}{\max(I(w, w_1), I(w, w_2))} \quad (4)$$

[3]The frequency based estimate for the expected frequency of a cooccurrence pair, assuming independent occurrence of the two words and using their individual frequencies, is $\frac{d}{N} f(w_1) f(w_2)$. As mentioned earlier, we use this estimate as representative for smoothing estimates of unobserved cooccurrences.

This way we get a uniform scale between 0 and 1, in which higher values reflect higher similarity. If both mutual information values are 0, then $sim_L(w_1, w_2, w)$ is defined to be 0. The *right context similarity*, $sim_R(w_1, w_2, w)$, is defined equivalently, for $I(w_1, w)$ and $I(w_2, w)$[4].

Using definition 4 for each word w in the lexicon, we get $2 \cdot l$ similarity values for w_1 and w_2, where l is the size of the lexicon. The general similarity between w_1 and w_2, termed $sim(w_1, w_2)$, is defined as a weighted average of these $2 \cdot l$ values. It is necessary to use some weighting mechanism, since small values of mutual information tend to be less significant and more vulnerable to noisy data. We found that the maximal value involved in computing the similarity relative to a specific word provides a useful weight for this word in computing the average. Thus, the weight for a specific left context similarity value, $W_L(w_1, w_2, w)$, is defined as:

$$W_L(w_1, w_2, w) = \max(I(w, w_1), I(w, w_2)) \quad (5)$$

(notice that this is the same as the denominator in definition 4). This definition provides intuitively appropriate weights, since we would like to give more weight to context words that have a large mutual information value with at least one of w_1 and w_2. The mutual information value with the other word may then be large, providing a strong "vote" for similarity, or may be small, providing a strong "vote" against similarity. The weight for a specific right context similarity value is defined equivalently. Using these weights, we get the weighted average in Figure 2 as the general definition of

[4]In the case of cooccurrence pairs, a word may be involved in two types of relations, being the left or right argument of the pair. The definitions can be easily adopted to cases in which there are more types of relations, such as provided by syntactic parsing.

$$sim(w_1, w_2) = \frac{\sum_{w \in lexicon} sim_L(w_1, w_2, w) \cdot W_L(w_1, w_2, w) + sim_R(w_1, w_2, w) \cdot W_R(w_1, w_2, w)}{\sum_{w \in lexicon} W_L(w_1, w_2, w) + W_R(w_1, w_2, w)} = \frac{\sum_{w \in lexicon} \min(I(w, w_1), I(w, w_2)) + \min(I(w_1, w), I(w_2, w))}{\sum_{w \in lexicon} \max(I(w, w_1), I(w, w_2)) + \max(I(w_1, w), I(w_2, w))} \quad (6)$$

Figure 2: The definition of the similarity metric.

Exhaustive Search		Approximation	
similar words	*sim*	*similar words*	*sim*
aspects	1.000	aspects	1.000
topics	0.100	topics	0.100
areas	0.088	areas	0.088
expert	0.079	expert	0.079
issues	0.076	issues	0.076
approaches	0.072	concerning	0.069

Table 3: The most similar words of *aspects:* heuristic and exhaustive search produce nearly the same results.

similarity[5].

The values produced by our metric have an intuitive interpretation, as denoting a "typical" ratio between the mutual information values of each of the two words with another third word. The metric is reflexive ($sim(w, w) = 1$), symmetric ($sim(w_1, w_2) = sim(w_2, w_1)$), but is not transitive (the values of $sim(w_1, w_2)$ and $sim(w_2, w_3)$ do not imply anything on the value of $sim(w_1, w_3)$). The left column of Table 3 lists the six most similar words to the word 'aspects' according to this metric, based on our corpus. More examples of similarity were shown in Figure 1.

4.1 An efficient search heuristic

The estimation method of section 3 requires that we identify the most similar words of a given word w. Doing this by computing the similarity between w and each word in the lexicon is computationally very expensive ($O(l^2)$, where l is the size of the lexicon, and $O(l^3)$ to do this in advance for all the words in the lexicon). To account for this problem we developed a simple heuristic that searches for words that are potentially similar to w, using thresholds on mutual information values and frequencies of cooccurrence pairs. The search is based on the property that when computing $sim(w_1, w_2)$, words that have high mutual information values with both w_1 and w_2 make the largest contributions to the value of the similarity measure. Also, high and reliable mutual information values are typically associated with relatively high frequencies of the involved cooccurrence pairs. We therefore search first for all the "strong neighbors" of w, which are defined as words whose cooccurrence with w has high mutual information and high frequency, and then search for all their "strong neighbors". The words found this way ("the strong neighbors of the strong neighbors of w") are considered as candidates for being similar words of w, and the similarity value with w is then computed only for these words. We thus get an approximation for the set of words that are most similar to w. For the example given in Table 3, the exhaustive method required 17 minutes of CPU time on a Sun 4 workstation, while the approximation required only 7 seconds. This was done using a data base of 1,377,653 cooccurrence pairs that were extracted from the corpus, along with their counts.

5 Evaluations

5.1 Word sense disambiguation in machine translation

The purpose of the first evaluation was to test whether the similarity based estimation method can enhance the performance of a disambiguation technique. Typically in a disambiguation task, different cooccurrences correspond to alternative interpretations of the ambiguous construct. It is therefore necessary that the probability estimates for the alternative cooccurrences will reflect the relative order between their true probabilities. However, a consistent bias in the estimate is usually not harmful, as it still preserves the correct relative order between the alternatives.

To carry out the evaluation, we implemented a variant of the disambiguation method of (Dagan et al., 1991), for sense disambiguation in machine translation. We term this method as *TWS*, for *Target Word Selection*. Consider for example the Hebrew phrase 'laxtom xoze shalom', which translates as 'to sign a peace treaty'. The word 'laxtom', however, is ambiguous, and can be translated to either 'sign' or 'seal'. To resolve the ambiguity, the

[5]The nominator in our metric resembles the similarity metric in (Hindle, 1990). We found, however, that the difference between the two metrics is important, because the denominator serves as a normalization factor.

	Precision	Applicability
TWS	85.5	64.3
Augmented TWS	83.6	79.6
Word Frequency	66.9	100

Table 4: Results of TWS, Augmented TWS and Word Frequency methods

TWS method first generates the alternative lexical cooccurrence patterns in the *target* language, that correspond to alternative selections of target words. Then, it prefers those target words that generate more frequent patterns. In our example, the word 'sign' is preferred upon the word 'seal', since the pattern 'to sign a treaty' is much more frequent than the pattern 'to seal a treaty'. Similarly, the word 'xoze' is translated to 'treaty' rather than 'contract', due to the high frequency of the pattern 'peace treaty'[6]. In our implementation, cooccurrence pairs were used instead of lexical cooccurrence within syntactic relations (as in the original work), to save the need of parsing the corpus.

We randomly selected from a software manual a set of 269 examples of ambiguous Hebrew words in translating Hebrew sentences to English. The expected success rate of random selection for these examples was 23%. The similarity based estimation method was used to estimate the expected frequency of unobserved cooccurrence pairs, in cases where none of the alternative pairs occurred in the corpus (each pair corresponds to an alternative target word). Using this method, which we term *Augmented TWS*, 41 additional cases were disambiguated, relative to the original method. We thus achieved an increase of about 15% in the applicability (coverage) of the TWS method, with a small decrease in the overall precision. The performance of the Augmented TWS method on these 41 examples was about 15% higher than that of a naive, *Word Frequency* method, which always selects the most frequent translation. It should be noted that the Word Frequency method is equivalent to using the frequency based estimate, in which higher word frequencies entail a higher estimate for the corresponding cooccurrence. The results of the experiment are summarized in Table 4.

5.2 A data recovery task

In the second evaluation, the estimation method had to distinguish between members of two sets of cooccurrence pairs, one of them containing pairs with relatively high probability and the other pairs with low probability. To a large extent, this task simulates a typical scenario in disambiguation, as demonstrated in the first evaluation.

Ideally, this evaluation should be carried out using a large set of held out data, which would provide good estimates for the true probabilities of the pairs in the test sets. The estimation method should then use a much smaller training corpus, in which none of the example pairs occur, and then should try to recover the probabilities that are known to us from the held out data. However, such a setting requires that the held out corpus would be several times larger than the training corpus, while the latter should be large enough for robust application of the estimation method. This was not feasible with the size of our corpus, and the rather noisy data we had.

To avoid this problem, we obtained the set of pairs with high probability from the training corpus, selecting pairs that occur at least 5 times. We then deleted these pairs from the data base that is used by the estimation method, forcing the method to recover their probabilities using the other pairs of the corpus. The second set, of pairs with low probability, was obtained by constructing pairs that do not occur in the corpus. The two sets, each of them containing 150 pairs, were constructed randomly and were restricted to words with individual frequencies between 500 and 2500. We term these two sets as the *occurring* and *non-occurring* sets.

The task of distinguishing between members of the two sets, without access to the deleted frequency information, is by no means trivial. Trying to use the individual word frequencies will result in performance close to that of using random selection. This is because the individual frequencies of all participating words are within the same range of values.

To address the task, we used the following procedure: The frequency of each cooccurrence pair was estimated using the similarity-based estimation method. If the estimated frequency was above 2.5 (which was set arbitrarily as the average of 5 and 0), the pair was recovered as a member of the *occurring* set. Otherwise, it was recovered as a member of the *non-occurring* set.

Out of the 150 pairs of the *occurring* set, our method correctly identified 119 (79%). For the *non-occurring* set, it correctly identified 126 pairs (84%). Thus, the method achieved an overall accuracy of 81.6%. Optimal tuning of the threshold, to a value of 2, improves the overall accuracy to 85%, where about 90% of the members of the *occurring* set and 80% of those in the *non-occurring*

[6] It should be emphasized that the TWS method uses only a *monolingual* target corpus, and not a bilingual corpus as in other methods ((Brown et al., 1991; Gale et al., 1992)). The alternative cooccurrence patterns in the target language, which correspond to the alternative translations of the ambiguous source words, are constructed using a bilingual lexicon.

set are identified correctly. This is contrasted with the optimal discrimination that could be achieved by frequency based estimation, which is 58%.

Figures 3 and 4 illustrate the results of the experiment. Figure 3 shows the distributions of the expected frequency of the pairs in the two sets, using similarity based and frequency based estimation. It clearly indicates that the similarity based method gives high estimates mainly to members of the *occurring* set and low estimates mainly to members of the *non-occurring* set. Frequency based estimation, on the other hand, makes a much poorer distinction between the two sets. Figure 4 plots the two types of estimation for pairs in the *occurring* set as a function of their true frequency in the corpus. It can be seen that while the frequency based estimates are always low (by construction) the similarity based estimates are in most cases closer to the true value.

6 Conclusions

In both evaluations, similarity based estimation performs better than frequency based estimation. This indicates that when trying to estimate cooccurrence probabilities, it is useful to consider the cooccurrence patterns of the specific words and not just their frequencies, as smoothing methods do. Comparing with class based models, our approach suggests the advantage of making the most specific analogies for each word, instead of making analogies with all members of a class, via general class parameters. This raises the question whether generalizations over word classes, which follow long traditions in semantic classification, indeed provide the best means for inferencing about properties of words.

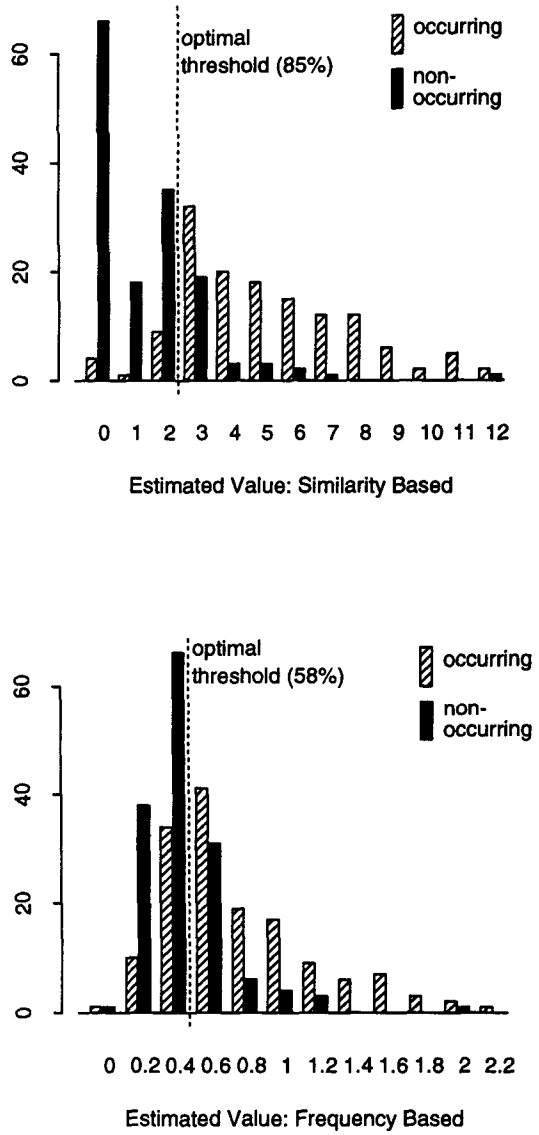

Figure 3: Frequency distributions of estimated frequency values for *occurring* and *non-occurring* sets.

Acknowledgements

We are grateful to Alon Itai for his help in initiating this research. We would like to thank Ken Church and David Lewis for their helpful comments on earlier drafts of this paper.

REFERENCES

Peter Brown, Vincent Della Pietra, Peter deSouza, Jenifer Lai, and Robert Mercer. Class-based n-gram models of natural language. *Computational Linguistics.* (To appear).

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1991. Word sense disambiguation using statistical methods. In *Proc. of the Annual Meeting of the ACL.*

Kenneth W. Church and William A. Gale. 1991. A comparison of the enhanced Good-Turing

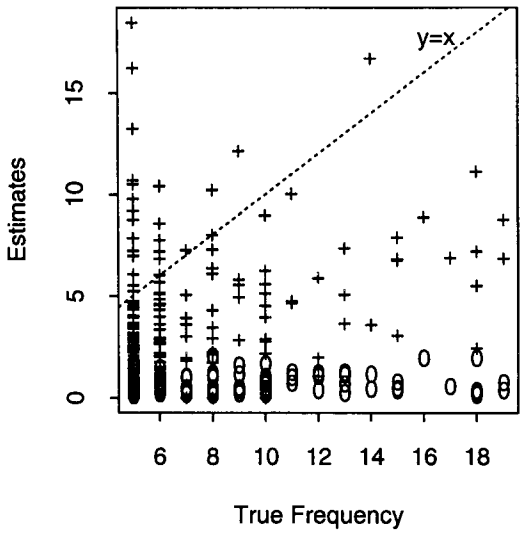

Figure 4: Similarity based estimation ('+') and frequency based estimation ('0') for the expected frequency of members of the *occurring* set, as a function of the true frequency.

and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Kenneth W. Church and Robert L. Mercer. 1992. Introduction to the special issue in computational linguistics using large corpora. *Computational Linguistics*. (In press).

Ido Dagan and Alon Itai. 1990. Automatic acquisition of constraints for the resolution of anaphora references and syntactic ambiguities. In *Proc. of COLING*.

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *Proc. of the Annual Meeting of the ACL*.

R. Fano. 1961. *Transmission of Information*. Cambridge,Mass:MIT Press.

William Gale, Kenneth Church, and David Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proc. of the International Conference on Theoretical and Methodolgical Issues in Machine Translation*.

I. J. Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264.

R. Grishman, L. Hirschman, and Ngo Thanh Nhan. 1986. Discovery procedures for sublanguage selectional patterns – initial experiments. *Computational Linguistics*, 12:205–214.

D. Hindle and M. Rooth. 1991. Structural ambiguity and lexical relations. In *Proc. of the Annual Meeting of the ACL*.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *Proc. of the Annual Meeting of the ACL*.

L. Hirschman. 1986. Discovering sublanguage structures. In R. Grishman and R. Kittredge, editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pages 211–234. Lawrence Erlbaum Associates.

F. Jelinek and R. Mercer. 1985. Probability distribution estimation from sparse data. *IBM Technical Disclosure Bulletin*, 28:2591–2594.

Frederick Jelinek. 1990. Self-organized language modeling for speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann Publishers, Inc., San Maeio, California.

Slava M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, speech, and Signal Processing*, 35(3):400–401.

Yoelle Maarek and Frank Smadja. 1989. Full text indexing based on lexical relations – An application: Software libraries. In *Proc. of SIGIR*.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proc. of the Annual Meeting of the ACL*.

Philip Resnik. 1992. Wordnet and distributional analysis: A class-based approach to lexical discovery. In *AAAI Workshop on Statistically-based Natural Language Processing Techniques*, July.

V. Sadler. 1989. *Working with analogical semantics: Disambiguation techniques in DLT*. Foris Publications.

Frank Smadja and Katheleen McKeown. 1990. Automatically extracting and representing collocations for language generation. In *Proc. of the Annual Meeting of the ACL*.