Appears in the proceedings of ACL-EACL '97

Similarity-Based Methods For Word Sense Disambiguation

Ido Dagan
Dept. of Mathematics and Computer Science
Bar Ilan University
Ramat Gan 52900, Israel
dagan@macs.biu.ac.il

Lillian Lee
Div. of Engineering and Applied Sciences
Harvard University
Cambridge, MA 01238, USA
llee@eecs.harvard.edu

Fernando Pereira
AT&T Labs – Research
600 Mountain Ave.
Murray Hill, NJ 07974, USA
pereira@research.att.com

Abstract

We compare four similarity-based estimation methods against back-off and maximum-likelihood estimation methods on a pseudo-word sense disambiguation task in which we controlled for both unigram and bigram frequency. The similarity-based methods perform up to 40% better on this particular task. We also conclude that events that occur only once in the training set have major impact on similarity-based estimates.

1 Introduction

The problem of data sparseness affects all statistical methods for natural language processing. Even large training sets tend to misrepresent low-probability events, since rare events may not appear in the training corpus at all.

We concentrate here on the problem of estimating the probability of *unseen* word pairs, that is, pairs that do not occur in the training set. Katz's back-off scheme (Katz, 1987), widely used in bigram language modeling, estimates the probability of an unseen bigram by utilizing unigram estimates. This has the undesirable result of assigning unseen bigrams the same probability if they are made up of unigrams of the same frequency.

Class-based methods (Brown et al., 1992; Pereira, Tishby, and Lee, 1993; Resnik, 1992) cluster words into classes of similar words, so that one can base the estimate of a word pair's probability on the averaged cooccurrence probability of the classes to which the two words belong. However, a word is therefore modeled by the average behavior of many words, which may cause the given word's idiosyncrasies to be ignored. For instance, the word "red" might well act like a generic color word in most cases, but it has distinctive cooccurrence patterns with respect to words like "apple," "banana," and so on.

We therefore consider similarity-based estimation schemes that do not require building general word classes. Instead, estimates for the most similar words to a word w are combined; the evidence provided by word w' is weighted by a function of its similarity to w. Dagan, Markus, and Markovitch (1993) propose such a scheme for predicting which unseen cooccurrences are more likely than others. However, their scheme does not assign probabilities. In what follows, we focus on probabilistic similarity-based estimation methods.

We compared several such methods, including that of Dagan, Pereira, and Lee (1994) and the *cooccurrence smoothing method* of Essen and Steinbiss (1992), against classical estimation methods, including that of Katz, in a decision task involving unseen pairs of direct objects and verbs, where unigram frequency was eliminated from being a factor. We found that all the similarity-based schemes performed almost 40% better than back-off, which is expected to yield about 50% accuracy in our experimental setting. Furthermore, a scheme based on the total divergence of empirical distributions to their average[1] yielded statistically significant improvement in error rate over cooccurrence smoothing.

We also investigated the effect of removing extremely low-frequency events from the training set. We found that, in contrast to back-off smoothing, where such events are often discarded from train-

[1]To the best of our knowledge, this is the first use of this particular distribution dissimilarity function in statistical language processing. The function itself is implicit in earlier work on distributional clustering (Pereira, Tishby, and Lee, 1993), has been used by Tishby (p.c.) in other distributional similarity work, and, as suggested by Yoav Freund (p.c.), it is related to results of Hoeffding (1965) on the probability that a given sample was drawn from a given joint distribution.

ing with little discernible effect, similarity-based smoothing methods suffer noticeable performance degradation when singletons (events that occur exactly once) are omitted.

2 Distributional Similarity Models

We wish to model conditional probability distributions arising from the coocurrence of linguistic objects, typically words, in certain configurations. We thus consider pairs $(w_1, w_2) \in V_1 \times V_2$ for appropriate sets V_1 and V_2, not necessarily disjoint. In what follows, we use subscript i for the i^{th} element of a pair; thus $P(w_2|w_1)$ is the conditional probability (or rather, some empirical estimate, the true probability being unknown) that a pair has second element w_2 given that its first element is w_1; and $P(w_1|w_2)$ denotes the probability estimate, according to the base language model, that w_1 is the first word of a pair given that the second word is w_2. $P(w)$ denotes the base estimate for the unigram probability of word w.

A similarity-based language model consists of three parts: a scheme for deciding which word pairs require a similarity-based estimate, a method for combining information from similar words, and, of course, a function measuring the similarity between words. We give the details of each of these three parts in the following three sections. We will only be concerned with similarity between words in V_1.

2.1 Discounting and Redistribution

Data sparseness makes the *maximum likelihood estimate (MLE)* for word pair probabilities unreliable. The MLE for the probability of a word pair (w_1, w_2), conditional on the appearance of word w_1, is simply

$$P_{ML}(w_2|w_1) = \frac{c(w_1, w_2)}{c(w_1)}, \tag{1}$$

where $c(w_1, w_2)$ is the frequency of (w_1, w_2) in the training corpus and $c(w_1)$ is the frequency of w_1. However, P_{ML} is zero for any unseen word pair, which leads to extremely inaccurate estimates for word pair probabilities.

Previous proposals for remedying the above problem (Good, 1953; Jelinek, Mercer, and Roukos, 1992; Katz, 1987; Church and Gale, 1991) adjust the MLE in so that the total probability of seen word pairs is less than one, leaving some probability mass to be redistributed among the unseen pairs. In general, the adjustment involves either *interpolation*, in which the MLE is used in linear combination with an estimator guaranteed to be nonzero for unseen word pairs, or *discounting*, in which a reduced MLE is used for seen word pairs, with the probability mass left over from this reduction used to model unseen pairs.

The discounting approach is the one adopted by Katz (1987):

$$\hat{P}(w_2|w_1) = \begin{cases} P_d(w_2|w_1) & c(w_1, w_2) > 0 \\ \alpha(w_1) P_r(w_2|w_1) & \text{o.w.} \end{cases}, \tag{2}$$

where P_d represents the Good-Turing discounted estimate (Katz, 1987) for seen word pairs, and P_r denotes the model for probability redistribution among the unseen word pairs. $\alpha(w_1)$ is a normalization factor.

Following Dagan, Pereira, and Lee (1994), we modify Katz's formulation by writing $P_r(w_2|w_1)$ instead $P(w_2)$, enabling us to use similarity-based estimates for unseen word pairs instead of basing the estimate for the pair on unigram frequency $P(w_2)$. Observe that similarity estimates are used for unseen word pairs only.

We next investigate estimates for $P_r(w_2|w_1)$ derived by averaging information from words that are distributionally similar to w_1.

2.2 Combining Evidence

Similarity-based models assume that if word w_1' is "similar" to word w_1, then w_1' can yield information about the probability of unseen word pairs involving w_1. We use a weighted average of the evidence provided by similar words, where the weight given to a particular word w_1' depends on its similarity to w_1.

More precisely, let $W(w_1, w_1')$ denote an increasing function of the similarity between w_1 and w_1', and let $\mathcal{S}(w_1)$ denote the set of words most similar to w_1. Then the general form of similarity model we consider is a W-weighted linear combination of predictions of similar words:

$$P_{\text{SIM}}(w_2|w_1) = \sum_{w_1' \in \mathcal{S}(w_1)} \frac{W(w_1, w_1')}{N(w_1)} P(w_2|w_1'), \tag{3}$$

where $N(w_1) = \sum_{w_1' \in \mathcal{S}(w_1)} W(w_1, w_1')$ is a normalization factor. According to this formula, w_2 is more likely to occur with w_1 if it tends to occur with the words that are most similar to w_1.

Considerable latitude is allowed in defining the set $\mathcal{S}(w_1)$, as is evidenced by previous work that can be put in the above form. Essen and Steinbiss (1992) and Karov and Edelman (1996) (implicitly) set $\mathcal{S}(w_1) = V_1$. However, it may be desirable to restrict $\mathcal{S}(w_1)$ in some fashion, especially if V_1 is large. For instance, Dagan, Pereira, and Lee (1994) use the closest k or fewer words w_1' such that the dissimilarity between w_1 and w_1' is less than a threshold value t; k and t are tuned experimentally.

Now, we could directly replace $P_r(w_2|w_1)$ in the back-off equation (2) with $P_{\mathrm{SIM}}(w_2|w_1)$. However, other variations are possible, such as interpolating with the unigram probability $P(w_2)$:

$$P_r(w_2|w_1) = \gamma P(w_2) + (1-\gamma) P_{\mathrm{SIM}}(w_2|w_1),$$

where γ is determined experimentally (Dagan, Pereira, and Lee, 1994). This represents, in effect, a linear combination of the similarity estimate and the back-off estimate: if $\gamma = 1$, then we have exactly Katz's back-off scheme. As we focus in this paper on alternatives for P_{SIM}, we will not consider this approach here; that is, for the rest of this paper, $P_r(w_2|w_1) = P_{\mathrm{SIM}}(w_2|w_1)$.

2.3 Measures of Similarity

We now consider several word similarity functions that can be derived automatically from the statistics of a training corpus, as opposed to functions derived from manually-constructed word classes (Resnik, 1992). All the similarity functions we describe below depend just on the base language model $P(\cdot|\cdot)$, not the discounted model $\hat{P}(\cdot|\cdot)$ from Section 2.1 above.

2.3.1 KL divergence

Kullback-Leibler (KL) divergence is a standard information-theoretic measure of the dissimilarity between two probability mass functions (Cover and Thomas, 1991). We can apply it to the conditional distribution $P(\cdot|w_1)$ induced by w_1 on words in V_2:

$$D(w_1 \| w_1') = \sum_{w_2} P(w_2|w_1) \log \frac{P(w_2|w_1)}{P(w_2|w_1')}. \qquad (4)$$

For $D(w_1\|w_1')$ to be defined it must be the case that $P(w_2|w_1') > 0$ whenever $P(w_2|w_1) > 0$. Unfortunately, this will not in general be the case for MLEs based on samples, so we would need smoothed estimates of $P(w_2|w_1')$ that redistribute some probability mass to zero-frequency events. However, using smoothed estimates for $P(w_2|w_1)$ as well requires a sum over all $w_2 \in V_2$, which is expensive for the large vocabularies under consideration. Given the smoothed denominator distribution, we set

$$W(w_1, w_1') = 10^{-\beta D(w_1 \| w_1')},$$

where β is a free parameter.

2.3.2 Total divergence to the average

A related measure is based on the total KL divergence to the average of the two distributions:

$$A(w_1, w_1') = D\left(w_1 \middle\| \frac{w_1 + w_1'}{2}\right) + D\left(w_1' \middle\| \frac{w_1 + w_1'}{2}\right), \qquad (5)$$

where $(w_1 + w_1')/2$ shorthand for the distribution

$$\frac{1}{2}(P(\cdot|w_1) + P(\cdot|w_1'))$$

Since $D(\cdot\|\cdot) \geq 0$, $A(w_1, w_1') \geq 0$. Furthermore, letting $p(w_2) = P(w_2|w_1)$, $p'(w_2) = P(w_2|w_1')$ and $\mathcal{C} = \{w_2 : p(w_2) > 0, p'(w_2) > 0\}$, it is straightforward to show by grouping terms appropriately that

$$\begin{aligned} A(w_1, w_1') = \; & \textstyle\sum_{w_2 \in \mathcal{C}} \Big\{ H(p(w_2) + p'(w_2)) \\ & \quad - H(p(w_2)) - H(p'(w_2)) \Big\} \\ + \; & 2 \log 2, \end{aligned}$$

where $H(x) = -x \log x$. Therefore, $A(w_1, w_1')$ is bounded, ranging between 0 and $2\log 2$, and smoothed estimates are not required because probability ratios are not involved. In addition, the calculation of $A(w_1, w_1')$ requires summing only over those w_2 for which $P(w_2|w_1)$ and $P(w_2|w_1')$ are both non-zero, which, for sparse data, makes the computation quite fast.

As in the KL divergence case, we set $W(w_1, w_1')$ to be $10^{-\beta A(w_1, w_1')}$.

2.3.3 L_1 norm

The L_1 *norm* is defined as

$$L(w_1, w_1') = \sum_{w_2} |P(w_2|w_1) - P(w_2|w_1')|. \qquad (6)$$

By grouping terms as before, we can express $L(w_1, w_1')$ in a form depending only on the "common" w_2:

$$\begin{aligned} L(w_1, w_1') \; = \; & 2 - \sum_{w_2 \in C} p(w_2) - \sum_{w_2 \in C} p'(w_2) \\ & + \sum_{w_2 \in C} |p(w_2) - p'(w_2)|. \end{aligned}$$

This last form makes it clear that $0 \leq L(w_1, w_1') \leq 2$, with equality if and only if there are no words w_2 such that both $P(w_2|w_1)$ and $P(w_2|w_1')$ are strictly positive.

Since we require a weighting scheme that is decreasing in L, we set

$$W(w_1, w_1') = (2 - L(w_1, w_1'))^{\beta}$$

with β again free.

2.3.4 Confusion probability

Essen and Steinbiss (1992) introduced *confusion probability* [2], which estimates the probability that word w_1' can be substituted for word w_1:

$$\begin{aligned} P_C(w_1'|w_1) &= W(w_1, w_1') \\ &= \sum_{w_2} \frac{P(w_1|w_2)P(w_1'|w_2)P(w_2)}{P(w_1)} \end{aligned}$$

Unlike the measures described above, w_1 may not necessarily be the "closest" word to itself, that is, there may exist a word w_1' such that $P_C(w_1'|w_1) > P_C(w_1|w_1)$.

The confusion probability can be computed from empirical estimates provided all unigram estimates are nonzero (as we assume throughout). In fact, the use of smoothed estimates like those of Katz's back-off scheme is problematic, because those estimates typically do not preserve consistency with respect to marginal estimates and Bayes's rule. However, using consistent estimates (such as the MLE), we can rewrite P_C as follows:

$$P_C(w_1'|w_1) = \sum_{w_2} \frac{P(w_2|w_1)}{P(w_2)} \cdot P(w_2|w_1')P(w_1').$$

This form reveals another important difference between the confusion probability and the functions D, A, and L described in the previous sections. Those functions rate w_1' as similar to w_1 if, roughly, $P(w_2|w_1')$ is high when $P(w_2|w_1)$ is. $P_C(w_1'|w_1)$, however, is greater for those w_1' for which $P(w_1', w_2)$ is large when $P(w_2|w_1)/P(w_2)$ is. When the ratio $P(w_2|w_1)/P(w_2)$ is large, we may think of w_2 as being exceptional, since if w_2 is infrequent, we do not expect $P(w_2|w_1)$ to be large.

2.3.5 Summary

Several features of the measures of similarity listed above are summarized in table 1. "Base LM constraints" are conditions that must be satisfied by the probability estimates of the base language model. The last column indicates whether the weight $W(w_1, w_1')$ associated with each similarity function depends on a parameter that needs to be tuned experimentally.

3 Experimental Results

We evaluated the similarity measures listed above on a word sense disambiguation task, in which each method is presented with a noun and two verbs, and decides which verb is more likely to have the noun as a direct object. Thus, we do not measure the absolute quality of the assignment of probabilities, as would be the case in a perplexity evaluation, but rather the relative quality. We are therefore able to ignore constant factors, and so we neither normalize the similarity measures nor calculate the denominator in equation (3).

3.1 Task: Pseudo-word Sense Disambiguation

In the usual word sense disambiguation problem, the method to be tested is presented with an ambiguous word in some context, and is asked to identify the correct sense of the word from the context. For example, a test instance might be the sentence fragment "robbed the bank"; the disambiguation method must decide whether "bank" refers to a river bank, a savings bank, or perhaps some other alternative.

While sense disambiguation is clearly an important task, it presents numerous experimental difficulties. First, the very notion of "sense" is not clearly defined; for instance, dictionaries may provide sense distinctions that are too fine or too coarse for the data at hand. Also, one needs to have training data for which the correct senses have been assigned, which can require considerable human effort.

To circumvent these and other difficulties, we set up a pseudo-word disambiguation experiment (Schütze, 1992; Gale, Church, and Yarowsky, 1992) the general format of which is as follows. We first construct a list of *pseudo-words*, each of which is the combination of two different words in V_2. Each word in V_2 contributes to exactly one pseudo-word. Then, we replace each w_2 in the test set with its corresponding pseudo-word. For example, if we choose to create a pseudo-word out of the words "make" and "take", we would change the test data like this:

make	plans	$\Rightarrow$	{make, take} plans
take	action	$\Rightarrow$	{make, take} action

The method being tested must choose between the two words that make up the pseudo-word.

3.2 Data

We used a statistical part-of-speech tagger (Church, 1988) and pattern matching and concordancing tools (due to David Yarowsky) to identify transitive main verbs and head nouns of the corresponding direct objects in 44 million words of 1988 Associated Press newswire. We selected the noun-verb pairs for the 1000 most frequent nouns in the corpus. These pairs are undoubtedly somewhat noisy given the errors

[2] Actually, they present two alternative definitions. We use their model 2-B, which they found yielded the best experimental results.

name	range	base LM constraints	tune?
D	$[0, \infty]$	$P(w_2\|w_1') \neq 0$ if $P(w_2\|w_1) \neq 0$	yes
A	$[0, 2\log 2]$	none	yes
L	$[0, 2]$	none	yes
P_{C}	$[0, \frac{1}{2}\max_{w_2} P(w_2)]$	Bayes consistency	no

Table 1: Summary of similarity function properties

inherent in the part-of-speech tagging and pattern matching.

We used 80%, or 587833, of the pairs so derived, for building base bigram language models, reserving 20% for testing purposes. As some, but not all, of the similarity measures require smoothed language models, we calculated both a Katz back-off language model ($P = \hat{P}$ (equation (2)), with $P_r(w_2|w_1) = P(w_2)$), and a maximum-likelihood model ($P = P_{\mathrm{ML}}$). Furthermore, we wished to investigate Katz's claim that one can delete *singletons*, word pairs that occur only once, from the training set without affecting model performance (Katz, 1987); our training set contained 82407 singletons. We therefore built four base language models, summarized in Table 2.

	with singletons (587833 pairs)	no singletons (505426 pairs)
MLE	MLE-1	MLE-o1
Katz	BO-1	BO-o1

Table 2: Base Language Models

Since we wished to test the effectiveness of using similarity for unseen word cooccurrences, we removed from the test set any verb-object pairs that occurred in the training set; this resulted in 17152 *unseen* pairs (some occurred multiple times). The unseen pairs were further divided into five equal-sized parts, T_1 through T_5, which formed the basis for fivefold cross-validation: in each of five runs, one of the T_i was used as a performance test set, with the other 4 sets combined into one set used for tuning parameters (if necessary) via a simple grid search. Finally, test pseudo-words were created from pairs of verbs with similar frequencies, so as to control for word frequency in the decision task. We use error rate as our performance metric, defined as

$$\frac{1}{N}(\text{\# of incorrect choices } + (\text{\# of ties})/2)$$

where N was the size of the test corpus. A tie occurs when the two words making up a pseudo-word are deemed equally likely.

3.3 Baseline Experiments

The performances of the four base language models are shown in table 3. MLE-1 and MLE-o1 both have error rates of exactly .5 because the test sets consist of unseen bigrams, which are all assigned a probability of 0 by maximum-likelihood estimates, and thus are all ties for this method. The back-off models BO-1 and BO-o1 also perform similarly.

	T_1	T_2	T_3	T_4	T_5
MLE-1	.5	.5	.5	.5	.5
MLE-o1	"	"	"	"	"
BO-1	0.517	0.520	0.512	0.513	0.516
BO-o1	0.517	0.520	0.512	0.513	0.516

Table 3: Base Language Model Error Rates

Since the back-off models consistently performed worse than the MLE models, we chose to use only the MLE models in our subsequent experiments. Therefore, we only ran comparisons between the measures that could utilize unsmoothed data, namely, the L_1 norm, $L(w_1, w_1')$; the total divergence to the average, $A(w_1, w_1')$; and the confusion probability, $P_{\mathrm{C}}(w_1'|w_1)$. [3] In the full paper, we give detailed examples showing the different neighborhoods induced by the different measures, which we omit here for reasons of space.

3.4 Performance of Similarity-Based Methods

Figure 1 shows the results on the five test sets, using MLE-1 as the base language model. The parameter β was always set to the optimal value for the corresponding training set. RAND, which is shown for comparison purposes, simply chooses the weights $W(w_1, w_1')$ randomly. $\mathcal{S}(w_1)$ was set equal to V_1 in all cases.

The similarity-based methods consistently outperform the MLE method (which, recall, always has an error rate of .5) and Katz's back-off method (which always had an error rate of about .51) by

[3] It should be noted, however, that on BO-1 data, KL-divergence performed slightly better than the L_1 norm.

a huge margin; therefore, we conclude that information from other word pairs is very useful for unseen pairs where unigram frequency is not informative. The similarity-based methods also do much better than RAND, which indicates that it is not enough to simply combine information from other words arbitrarily: it is quite important to take word similarity into account. In all cases, A edged out the other methods. The average improvement in using A instead of P_{C} is .0082; this difference is significant to the .1 level ($p < .085$), according to the paired t-test.

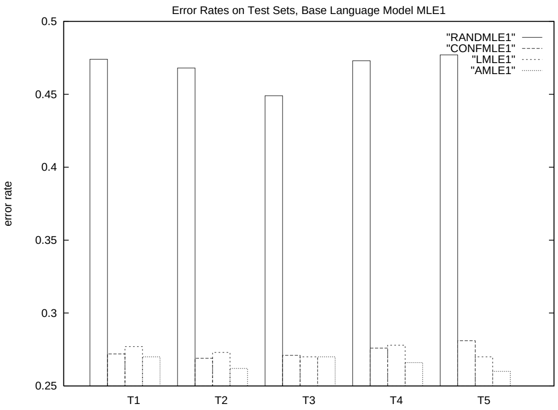

Figure 1: Error rates for each test set, where the base language model was MLE-1. The methods, going from left to right, are RAND , P_{C}, L, and A. The performances shown are for settings of β that were optimal for the corresponding training set. β ranged from 4.0 to 4.5 for L and from 10 to 13 for A.

The results for the MLE-o1 case are depicted in figure 2. Again, we see the similarity-based methods achieving far lower error rates than the MLE, back-off, and RAND methods, and again, A always performed the best. However, with singletons omitted the difference between A and P_{C} is even greater, the average difference being .024, which is significant to the .01 level (paired t-test).

An important observation is that all methods, including RAND, were much more effective if singletons were included in the base language model; thus, in the case of unseen word pairs, Katz's claim that singletons can be safely ignored in the back-off model does not hold for similarity-based models.

4 Conclusions

Similarity-based language models provide an appealing approach for dealing with data sparseness. We have described and compared the performance

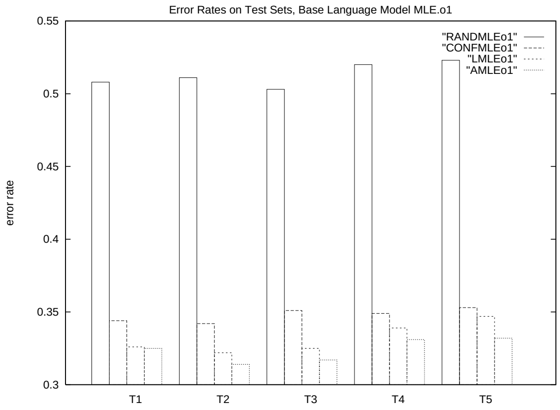

Figure 2: Error rates for each test set, where the base language model was MLE-o1. β ranged from 6 to 11 for L and from 21 to 22 for A.

of four such models against two classical estimation methods, the MLE method and Katz's back-off scheme, on a pseudo-word disambiguation task. We observed that the similarity-based methods perform much better on unseen word pairs, with the measure based on the KL divergence to the average, being the best overall.

We also investigated Katz's claim that one can discard singletons in the training data, resulting in a more compact language model, without significant loss of performance. Our results indicate that for similarity-based language modeling, singletons are quite important; their omission leads to significant degradation of performance.

Acknowledgments

We thank Hiyan Alshawi, Joshua Goodman, Rebecca Hwa, Stuart Shieber, and Yoram Singer for many helpful comments and discussions. Part of this work was done while the first and second authors were visiting AT&T Labs. This material is based upon work supported in part by the National Science Foundation under Grant No. IRI-9350192. The second author also gratefully acknowledges support from a National Science Foundation Graduate Fellowship and an AT&T GRPW/ALFP grant.

References

[Brown et al.1992] Brown, Peter F., Vincent J. DellaPietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, December.

[Church1988] Church, Kenneth. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143.

[Church and Gale1991] Church, Kenneth W. and William A. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54.

[Cover and Thomas1991] Cover, Thomas M. and Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley.

[Dagan, Pereira, and Lee1994] Dagan, Ido, Fernando Pereira, and Lillian Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd Annual Meeting of the ACL*, pages 272–278, Las Cruces, NM.

[Essen and Steinbiss1992] Essen, Ute and Volker Steinbiss. 1992. Co-occurrence smoothing for stochastic language modeling. In *Proceedings of ICASSP*, volume 1, pages 161–164.

[Gale, Church, and Yarowsky1992] Gale, William, Kenneth Church, and David Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *Working Notes, AAAI Fall Symposium Series, Probabilistic Approaches to Natural Language*, pages 54–60.

[Good1953] Good, I.J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3 and 4):237–264.

[Hoeffding1965] Hoeffding, Wassily. 1965. Asymptotically optimal tests for multinomial distributions. *Annals of Mathematical Statistics*, pages 369–401.

[Jelinek, Mercer, and Roukos1992] Jelinek, Frederick, Robert L. Mercer, and Salim Roukos. 1992. Principles of lexical language modeling for speech recognition. In Sadaoki Furui and M. Mohan Sondhi, editors, *Advances in Speech Signal Processing*. Mercer Dekker, Inc., pages 651–699.

[Karov and Edelman1996] Karov, Yael and Shimon Edelman. 1996. Learning similarity-based word sense disambiguation from sparse data. In *4rth Workshop on Very Large Corpora*.

[Katz1987] Katz, Slava M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3):400–401, March.

[Pereira, Tishby, and Lee1993] Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 183–190, Columbus, OH.

[Resnik1992] Resnik, Philip. 1992. Wordnet and distributional analysis: A class-based approach to lexical discovery. *AAAI Workshop on Statistically-based Natural Language Processing Techniques*, pages 56–64, July.

[Schütze1992] Schütze, Hinrich. 1992. Context space. In *Working Notes, AAAI Fall Symposium on Probabilistic Approaches to Natural Language*.