On the Effectiveness of the Skew Divergence for Statistical Language Analysis*

Lillian Lee
Department of Computer Science
Cornell University, Ithaca, NY 14853 USA
llee@cs.cornell.edu

Abstract

Estimating word co-occurrence probabilities is a problem underlying many applications in statistical natural language processing. Distance-weighted (or similarity-weighted) averaging has been shown to be a promising approach to the analysis of novel co-occurrences. Many measures of distributional similarity have been proposed for use in the distance-weighted averaging framework; here, we empirically study their stability properties, finding that similarity-based estimation appears to make more efficient use of more reliable portions of the training data. We also investigate properties of the skew divergence, a weighted version of the Kullback-Leibler (KL) divergence; our results indicate that the skew divergence yields better results than the KL divergence even when the KL divergence is applied to more sophisticated probability estimates.

1 INTRODUCTION

Estimating the probability of *co-occurrences* of linguistic objects is a fundamental tool in statistical approaches to natural language processing. For example, consider the following speech understanding scenario. Two similar-sounding transcription hypotheses for a given utterance (presumably regarding an AI system) have been produced:

1. This machine understands you like your mother.
2. This machine understands your lie cured mother.

Although both alternatives are grammatical, it is clear that the first sequence of words is more likely and so should be the preferred analysis in the absence of any further information. Also, to interpret sentence 1, which is ambiguous, we need to determine whether it is more likely that the object of the verb "understands" is "you"(i.e., the machine understands you as well as your mother does) or "you like your mother" (that is, the machine knows that you like your mother). Note that we desire probability estimates, rather than decisions as to which alternative is "correct", because there are situations in which the second sentence makes more sense in context; indeed, the standard speech recognition architecture weighs these estimates against evidence from other knowledge sources, such as the system's acoustic model.

The general problem we consider here is the estimation of co-occurrence probabilities based solely on the frequencies of the co-occurrences themselves; this is the setting in which most state-of-the-art speech recognition systems are trained (Jelinek, 1997). A major challenge is estimating the probability of *novel* (previously unseen) co-occurrences, which in the natural language domain can make up a large percentage of test data even when huge training sets are employed (Brown et al., 1992; Rosenfeld, 1996).

A standard, state-of-the-art approach is to "back off" to the unigram probability when a new co-occurrence pair not found in the training data is observed: $\widehat{P}(y|x) \propto P(y)$ (Katz, 1987); Chen and Goodman (1996) show that this is one of the best methods for estimating co-occurrence pair probabilities. However, an intuitively appealing alternative is to form an estimate based on the frequencies of *similar* co-occurrences (Saul and Pereira, 1997; Hofmann and Puzicha, 1998; Dagan, Lee, and Pereira, 1999). This is, of course, the idea underlying a large body of work that is variously termed nearest-neighbor, case-based, memory-based, instance-based, and lazy learning, among other designations (Cover and Hart, 1967; Stanfill and Waltz, 1986; Aha, Kibler, and Albert, 1991; Atkeson, Moore, and Schaal, 1997).

*Appears in *Artificial Intelligence and Statistics 2001*, pp. 65-72.

Memory-based methods for classification have been successfully applied to a variety of language understanding tasks. Typically, similarity between objects is determined by the similarity of their corresponding feature vectors, where the features are symbolic (Stanfill, 1987; Cardie, 1993; Ng and Lee, 1996; Daelemans, van den Bosch, and Zavrel, 1999). In contrast, in our setting, the "features" are numeric co-occurrence frequencies. Also, memory-based work has concentrated on supervised learning, in which class labels are provided in the training data; however, co-occurrence probability estimation must be unsupervised because the true probability distribution of linguistic objects is not known. In short, our focus is on *distributional similarity*: determining the similarity between two empirically-determined conditional distributions $P(Y|x)$ and $P(Y|x')$.

In previous work (Dagan, Lee, and Pereira, 1999), we have demonstrated that using similarity-weighted averaging has the potential to produce high-quality probability estimates for low-frequency co-occurrences, yielding improvements on standard performance metrics of up to 20% with respect to back-off. Later work (Lee, 1999) compared the behavior of several measures of distributional similarity and introduced the *skew divergence*, an approximation to the Kullback-Leibler (KL) divergence, itself a classic measure of the distance between distributions (Cover and Thomas, 1991).

In this paper, we study the stability of similarity measures when trained on frequency-filtered data, finding that in comparison to Katz's back-off, similarity-based estimation is far more robust. This indicates that similarity functions may make more efficient use of arguably more reliable sections of the training data.

Also, we further study the skew divergence. We find that roughly speaking, the skew divergences yielding the best results are those that approximate the KL divergence more closely. This leads to the question of whether the good performance is due simply to "proximity" to the KL divergence, which, for technical reasons, cannot be directly applied to non-smoothed distributions. That is, we ask whether it is better to approximate the KL divergence, or to use the KL divergence itself on smoothed distributions. Our results indicate that even if we use the KL divergence on distributions derived via sophisticated estimation methods, the skew divergence achieves better results.

2 DISTRIBUTIONAL SIMILARITY

For two objects x and x' co-occurring with objects drawn from a finite set $\mathcal{Y}$, we seek to calculate the similarity (or distance) between $q(Y) \triangleq P(Y|x)$ and $r(Y) \triangleq P(Y|x')$, where these distributions are estimated from a training corpus. For example, we can infer that the word "business" might serve as a good proxy for the word "company" because they both frequently occur as the object of verbs like "acquire", but rarely as objects of the verb "defenestrate".

There are many possible similarity (or distance) functions. In section 2.1 we briefly introduce a few of the most commonly-used measures, listed in Table 1. Section 2.2 presents examples of similar words computed using these functions.

2.1 FUNCTIONS

The *Kullback-Leibler (KL) divergence* $D(q||r)$ is a classic measure of the "distance" between two probability mass functions (Cover and Thomas, 1991). Unfortunately, it is undefined if there exists a $y \in \mathcal{Y}$ such that $q(y) > 0$ but $r(y) = 0$. This property makes it unsuitable for distributions derived via maximum-likelihood estimates, which assign a probability of zero to co-occurrences not appearing in the training data. Unfortunately, in natural language tasks, such *unseen* co-occurrences are very common in application (test) data (Brown et al., 1992).

One option is to employ *smoothed* estimates so that $r(y)$ is non-zero for all y, as was done in our previous work (Dagan, Lee, and Pereira, 1999). Another alternative is to use approximations of the KL divergence that do not require q to be absolutely continuous with respect to r. Two such functions are the *Jensen-Shannon divergence*[1] (J. Lin, 1991) and the *skew divergence* (Lee, 1999). The Jensen-Shannon divergence, which is symmetric, considers the KL divergence between q, r, and the *average* of q and r, under the assumption that if q and r are similar to each other, they should both be "close" to their average. The asymmetric skew divergence, on the other hand, simply smooths one of the distributions by mixing it, to a degree determined by the parameter α, with the other distribution (observe from Table 1 that at $\alpha = 1$, the approximation is exact).

Another way to measure distributional similarity is to treat the distributions as vectors and apply geometrically-motivated functions. These include the *Euclidean* distance, the *cosine*, and the L_1 (or Manhattan) distance.

We also include in our study two functions of a somewhat different flavor that have been previously used in language processing tasks. The *confusion probability*, which estimates the substitutability of two given words (Sugawara et al., 1985; Essen and Steinbiss, 1992;

[1] The Jensen-Shannon divergence in Table 1 is a special case of the function defined by J. Lin (1991).

Table 1: Similarity functions for probability distributions. The function $\text{avg}(q,r)$ in the Jensen-Shannon divergence is the averaged distribution $(q(y)+r(y))/2$. The skew divergence constant α lies in the range $[0,1]$.

KL DIVERGENCE	$D(q \parallel r)$	=	$\sum_y q(y)(\log q(y) - \log r(y))$
JENSEN-SHANNON	$\text{JS}(q,r)$	=	$\frac{1}{2}\left[D\left(q \,\middle\|\, \text{avg}(q,r)\right) + D\left(r \,\middle\|\, \text{avg}(q,r)\right)\right]$
SKEW DIVERGENCE	$s_\alpha(q,r)$	=	$D\left(r \parallel \alpha q + (1-\alpha)r\right)$
EUCLIDEAN	$\text{euc}(q,r)$	=	$\left(\sum_y \left(q(y)-r(y)\right)^2\right)^{1/2}$
COSINE	$\cos(q,r)$	=	$\sum_y q(y)r(y) \Big/ \sqrt{\sum_y q(y)^2 \sum_y r(y)^2}$
L_1	$L_1(q,r)$	=	$\sum_y \lvert q(y)-r(y)\rvert$
CONFUSION	$\text{conf}(q,r,P(x'))$	=	$P(x')\sum_y q(y)r(y)/P(y)$
TAU	$\tau(q,r)$	=	$\sum_{y_1,y_2} \text{sign}\left[(q(y_1)-q(y_2))(r(y_1)-r(y_2))\right] \Big/ \left(2\binom{\lvert V\rvert}{2}\right)$

Table 2: Nearest neighbors to the word "company", most similar first. Italics designate words appearing in only one column. The words "government" and "organization" have been abbreviated.

	SKEW ($\alpha = .99$)	J.-S.	L_1	COSINE	CONFUSION	TAU	EUCLIDEAN
1	AIRLINE	BUSINESS	BUSINESS	*talk*	GOVT.	YEAR	CITY
2	BUSINESS	AIRLINE	AIRLINE	*hostage*	YEAR	STATE	AIRLINE
3	BANK	FIRM	STATE	*primary*	PEOPLE	PEOPLE	INDUSTRY
4	AGENCY	BANK	BANK	*referendum*	*percent*	GOVT.	PROGRAM
5	FIRM	STATE	FIRM	*lead*	*syndrome*	GROUP	ORG.
6	*department*	AGENCY	AGENCY	*hearing*	BUSINESS	*country*	BANK
7	*manufacturer*	GROUP	GOVT.	*stake*	TODAY	BUSINESS	SYSTEM
8	*network*	GOVT.	CITY	*discussion*	FIRM	PROGRAM	TODAY
9	INDUSTRY	CITY	GROUP	*post*	MEETING	TODAY	*series*
10	GOVT.	INDUSTRY	ORG.	MEETING	*stock*	SYSTEM	*portion*

Grishman and Sterling, 1993), is based not only on the conditional distributions $P(Y|x)$ and $P(Y|x')$ but also on marginal probabilities. The *tau coefficient,*[2] a statistical measure of the association between random variables (Liebetrau, 1983), is based on probability rankings rather than the actual probability values: roughly speaking, $\tau(q,r)$ is larger if there are few pairs (y_1,y_2) such that $q(y_1) > q(y_2)$ but $r(y_1) < r(y_2)$, or vice versa. Hatzivassiloglou (1996) applied this measure to the task of grouping related adjectives.

2.2 EXAMPLES OF SIMILAR WORDS

In order to provide some intuition about the measures described in the previous section, we now present an example of similar words derived by each function.

[2]There are actually three versions of the tau coefficient; we are using τ_a. See Liebetrau (1983) for discussion of this issue.

We computed similarity between nouns based on their occurrences as the heads of direct objects of verbs in a newswire corpus (details of the training set are given in section 3.2 below). That is, the training data consisted of noun-verb pairs such as $(company, acquire)$; we used maximum-likelihood estimation for the conditional distributions $P(Y|x)$, e.g.

$$P(acquire|company) = \frac{\#(company, acquire)}{\#(company)},$$

where $\#(\cdot)$ denotes frequency in the corpus. Then, for each noun x', we computed the similarity of $P(Y|x')$ to $P(Y|\,company)$. We note that no linguistic knowledge was employed other than the fact that certain nouns occurred as the objects of certain verbs.

Table 2 shows the ten nearest neighbors to the word "company" according to each of the measures described above (except the KL divergence, which, as mentioned earlier, cannot be applied to maximum-

likelihood estimates). The left-to-right sequence of the table foreshadows the functions' relative empirical performances in best-first order; these results are presented in section 3.3.

We first observe that many of the nearest neighbors, such as "business" and "firm", indeed seem semantically quite similar to the word "company". The L_1 distance and Jensen-Shannon divergence appear highly correlated.[3] But there is noticeable variation between the lists; for instance, no word is in the top ten for every metric, although "business" and "government" appear in five of the seven rankings. To highlight this variability, we have italicized all words that are in the top ten according to one similarity function but not considered a nearest neighbor by any of the others. Evidently, the cosine metric is strikingly different from the other measures, perhaps because of the length normalization it incorporates (note in Table 1 that otherwise the cosine looks somewhat similar to the confusion probability). Also, four of the other six lists contain novel words as well, with the two asymmetric functions (the confusion probability and the skew divergence) exhibiting a higher percentage of unique neighbors.

We thus see that the various functions can exhibit qualitatively different behaviors. In the next section, we evaluate their performances quantitatively.

3 EXPERIMENTS

3.1 METHODOLOGY

Our ultimate goal is the accurate estimation of co-occurrence probabilities. It would therefore seem that we should directly evaluate the probabilities assigned by similarity-based models. However, we then run into a methodological problem: how should we define the similarity-based models? An attractive paradigm, though certainly not the only possibility, is the *distance-weighted averaging* framework, where the relative frequencies of the nearest neighbors are averaged together after being weighted by their similarity to the target (Atkeson, Moore, and Schaal, 1997). Unfortunately, even if we settle on distance-weighted averaging, we must choose among many possible weight functions. The difficulty is that we want our stability analyses to be independent of these various decisions in order to factor out the choice of metric from the choice of model, since for many similarity measures it is not at all clear what the optimal weight function is. (In previous work, we have seen that different weights for the same similarity measure can lead to substantial decreases in performance (Dagan, Lee, and Pereira, 1999).)

We therefore employed the following experimental framework. Similarity functions are compared by examining their ability to choose the more likely of two test co-occurrences (x, y_1) and (x, y_2). The decision is made by taking a majority vote of the k nearest neighbors according to the measure under consideration, where a neighbor x' *votes* for (x, y_1) if $P(y_1|x') > P(y_2|x')$, and vice versa. By plotting performance as a function of the number of most similar neighbors considered, one can compare the *rankings* induced by different similarity functions independently of the numerical values they assign: we should clearly prefer function f over function g if, for all k, a higher percentage of f's k most similar words vote correctly. This methodology allows us to compare functions directly without needing to consider issues such as weighting functions, number of nearest neighbors selected, etc. In practice, of course, these factors are important. However, it is desirable to concentrate our efforts on finding weight functions for inherently better similarity functions, so it makes sense to first identify which similarity measures are the most promising before considering what the best weighting scheme is.

3.2 INITIAL DATA AND PERFORMANCE METRICS

The datasets were constructed from a collection of over 730,000 verb-noun co-occurrence pairs from the Associated Press 1988 newswire involving the 1000 most frequent nouns. These were extracted via part-of-speech tagging and pattern matching tools due to Church (1988) and David Yarowsky. We randomly selected 80% to serve as training data. From the remaining 20%, we created five equal-sized disjoint sets of co-occurrences not also appearing in the training corpus (recall that our goal is to model unseen co-occurrences). Finally, these five sets were transformed into test sets by replacing each pair (x, y_1) with the test instance $((x, y_1), (x, y_2))$, where y_2 was randomly chosen among those verbs with approximately the same frequency as y_1.[4]

In order to evaluate performance, we calculated the *error rate*, defined as

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

where T is the size of the test set. Ties — instances in

[3] Indeed, the L_1 distance bounds the Jensen-Shannon divergence (J. Lin, 1991).

[4] Other ways to create test alternatives include constructing instances in which two pairs (x, y_1) and (x, y_2) were both unseen in the training data but occur in the test data; however, this narrows the size of the test corpus considerably.

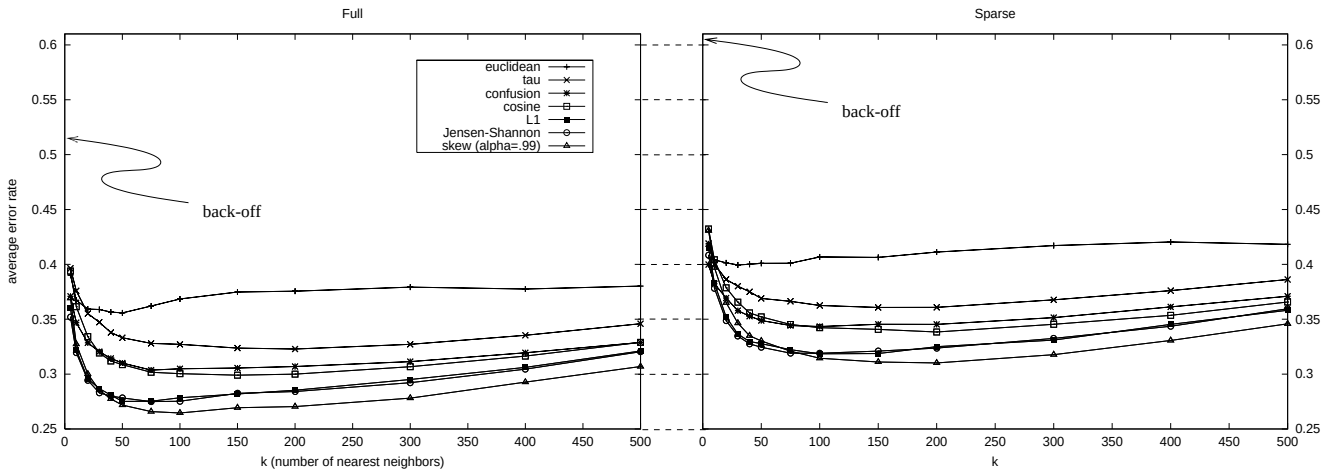

Figure 1: Average error rates for similarity measures as a function of k, the number of nearest neighbors. Left: trained on `full`. Right: trained on `sparse`. The key is ordered by performance.

which both alternatives are deemed equally likely — are treated as half-mistakes.

Each of the performance graphs depicts the average error rate over the five test sets (standard deviations were fairly small and so error bars are omitted for clarity). Unless otherwise specified, maximum-likelihood estimates were used for the base probabilities that serve as the input to back-off and the similarity measures. By our reasoning above, the "best" similarity measures should achieve the smallest error rates, and achieve these as early as possible.

3.3 EXPERIMENT 1: DEPENDENCE ON LOW-FREQUENCY EVENTS

In our first experiment, we examined the performance of similarity metrics on frequency-filtered training data. We are interested in the relative dependence of similarity measures on low-frequency events, since such events may constitute unreliable evidence. Indeed, Katz (1987) suggested that deleting singletons – events occurring only once – from the data can create more compact training sets without affecting the quality of statistical estimates. To study this issue, we constructed two training sets, `full` and `sparse`, where `full` was the training corpus described in the previous section. The `sparse` set was created from `full` by omitting all co-occurrence pairs appearing only once; this resulted in a 15% reduction in size.

As a baseline, we computed the performance of back-off, which does not use distributional similarity information. Recall that when presented with two unseen co-occurrences (x, y_1) and (x, y_2), back-off will prefer the alternative for which y_i is more frequent in the training corpus. When trained on `full`, back-off's average error rate was 51.5%: by construction of the test sets, back-off is reduced to random guessing. With `sparse` as the training set, the average error rate rose to 60.5%, for a difference of 9 percentage points.

Figure 1 shows the performance of the similarity measures for both training sets. Despite the large drop in training set size, the general shapes of the curves and relative performance orderings[5] are remarkably stable across the two training corpora, with the skew divergence yielding the best results.[6] The major changes are that the best achievable error rates rise — the average increase is 4.2 percentage points — and the minima are generally achieved at higher values of k for `sparse`, indicating that it is more difficult to select good nearest neighbors when singleton co-occurrences are missing. On the other hand, the performance degradation that the similarity measures suffer is less than half that for the baseline back-off. Thus, we conclude that in comparison to back-off, similarity measures do not depend as much on information carried by singleton events. This suggests that similarity-based estimation takes greater advantage of higher-frequency events, which may be more reliable information sources.

3.4 EXPERIMENT 2: SKEW VALUES

Given that a skew divergence achieved the best error rates in the previous experiment, we now study the family of skew divergences more carefully. Recall the

[5] A hypothesis as to why the functions seem to "clump" appears in our previous work (Lee, 1999).

[6] By the paired t-test, the differences between the skew divergence and the L_1 distance are significant at the .01 level for $k \geq 75$ (`full`) and $k \geq 200$ (`sparse`).

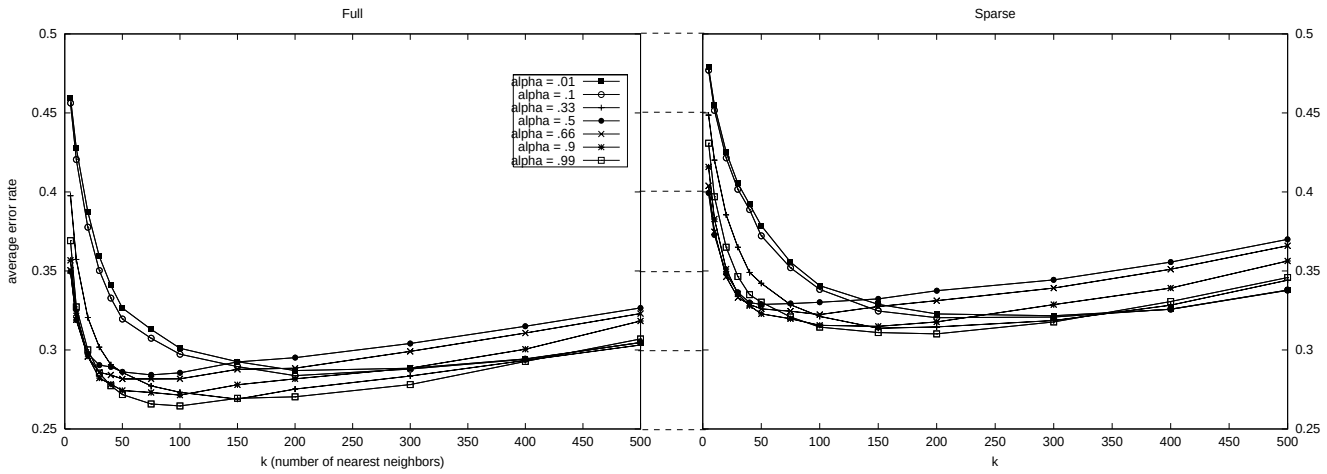

Figure 2: Average error rates for skew divergences trained on `full` (left) and `sparse` (right). The key is ordered by increasing α.

definition:

$$s_\alpha(q, r) = D\left(r \parallel \alpha q + (1-\alpha)r\right) \ ,$$

where α controls the degree to which the function approximates $D(r||q)$. In application, for a given test instance $((x, y_1), (x, y_2))$, we use the neighbors x' such that $s_\alpha(P(Y|x), P(Y|x'))$ is smallest — since we are using a similarity-based estimate for the distribution of verbs conditioned on object x, we are evidently considering the maximum-likelihood estimate to be unreliable, and hence it is the empirical $P(Y|x)$ that needs to be smoothed.[7]

Since the KL divergence is a standard and well-motivated distributional distance measure and α's role appears to be simply to guarantee that the skew divergence is always defined, two natural questions arise. First, does increasing α, thereby bringing the skew divergence closer to the KL divergence, always yield better results? Second, given that $(1-\alpha)r$ serves to smooth q, should the proper value of $(1-\alpha)$ depend on some measurement of the sparseness of the data?

To investigate these issues, we examined how varying the value of α changed the performance of the skew divergence for both of our training sets. The results are shown in figure 2. Again, we see that the shapes and relative orderings of the performance curves are preserved across training sets, with the error rates rising and the minima shifting to the right for `sparse`. And again, the skew divergences as a family are also less affected by frequency filtering than the baseline, back-off.

[7]Indeed, in our experiments, approximating $D(q||r)$ did not perform as well.

However, the role that α plays is not entirely clear. The highest value yielded the best performance and very small values resulted in the worst error rates, as one might expect; but the relationship between error rate and α for intermediate settings is not so simple. In particular, it appears that the skew divergence for $\alpha = 0.33$ does a relatively poor job of selecting good neighbors for small values of k, but the words ranked between 50 and 100 are very good predictors; as a consequence, it achieves lower best-case error rates than for the larger values $\alpha = 0.5$ and $\alpha = 0.66$.

In answer to the second question, Figure 2 shows that the best setting of α does not vary when we use `sparse` for training instead of `full`, which we may take as a stability result regarding high values of α. This finding may indicate that it is always better to choose a high setting of this parameter, thus obviating the need for extra training data to tune it; of course, further experiments are necessary to validate this claim.

3.5 EXPERIMENT 3: COMPARING SKEW AND KL DIVERGENCES

The previous experiment demonstrated that larger values of α are to be preferred; that is, the best results come about when we choose a skew divergence that is very close to the KL divergence D. It might therefore seem that the optimal thing to do is to actually use the KL divergence itself, rather than approximate versions.

As mentioned in section 2.1, we cannot apply the KL divergence directly to the maximum-likelihood estimates used as the input to the other similarity functions in experiments 1 and 2. However, we can first

apply some other estimation technique that produces *smoothed* distributions (i.e., where $P(y|x) > 0$ for all y), and use these as input to D. Then, the question is whether using the KL divergence on smoothed distributions can provide better results than using an approximation of the KL divergence on non-smoothed input. To investigate this issue, we looked at computing the KL divergence between probability distributions estimated using back-off, which we recall has been described as one of the best techniques for estimating the probability of co-occurrence pairs (Chen and Goodman, 1996).

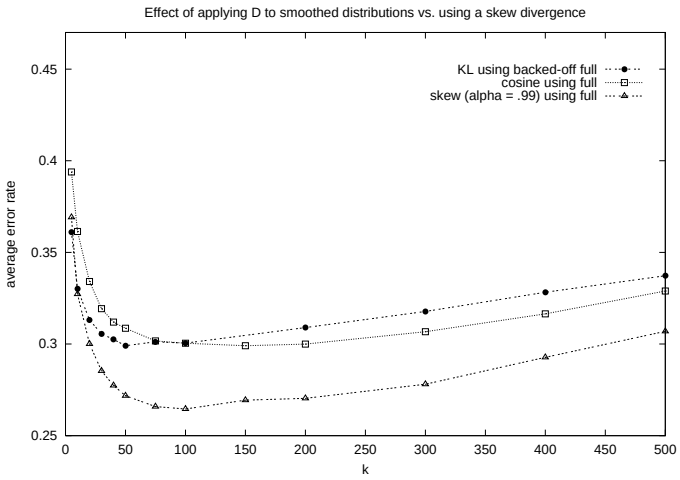

Figure 3: Average error rates for the KL divergence, skew divergence, and cosine trained on `full`.

Figure 3 shows that even when a state-of-the-art estimation method is used in conjunction with the KL divergence to calculate distributional similarity[8], it is better to use the skew divergence. Indeed, we see that D's error rate is on par with that of the cosine metric. Thus, it appears that highly sophisticated smoothing techniques would be needed in order to use the KL divergence effectively, so much so that the skew divergence may be of greater utility than the function it is supposed to approximate.

4 RELATED WORK

The distributional similarity functions we studied here have been applied to other natural language processing tasks. For example, McCarthy (2000) used the skew divergence to analyze verb arguments. Lapata (2000) looked at using similarity-based estimation to analyze nominalizations. Interestingly, she found that although the Jensen-Shannon divergence was very good for predicting object relations, the confusion probability was superior for subject relations. We intend to investigate this issue in future work.

The computation of distributional word similarity has also been proposed as a way to automatically construct thesauri; see Grefenstette (1994), Hatzivassiloglou (1996), D. Lin (1998) , and Caraballo (1999) for some recent examples.

This paper has considered distributional similarity in a nearest-neighbor, locally-weighted framework. However, an alternative is to build a *cluster*-based probability model that groups co-occurrences into global classes; examples in the language processing literature include the work of Brown et al. (1992), Pereira, Tishby, and Lee (1993), Kneser and Ney (1993), and Hofmann and Puzicha (1998). Rooth et al. (1999) use such clustering techniques to learn subcategorization frames and other lexical information. Recent work has attempted a comparison of the nearest-neighbor and clustering paradigms (Lee and Pereira, 1999).

Finally, we note that stability issues are clearly also relevant to memory-based classification, i.e., in *supervised* settings. Daelemans, van den Bosch, and Zavrel (1999) describe a suite of experiments exploring the effects on classification error of editing the training data in various ways.

Acknowledgements

We thank Fernando Pereira for access to computational resources at AT&T – Research. Part of this work was completed while visiting Harvard University; thanks to Stuart Shieber for this opportunity. The opening example is a modified version of an example also due to Stuart Shieber. The third experiment was motivated by an anonymous reviewer's comment on Lee (1999). Finally, we thank the anonymous reviewers for their helpful comments. This material is based upon work supported in part by the National Science Foundation under Grant No. IRI9712068.

References

Aha, David W., Dennis Kibler, and Marc K. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Atkeson, Christopher G., Andrew W. Moore, and Stefan Schaal. 1997. Locally weighted learning. *Artificial Intelligence Review*, 11(1):11–73.

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.

Caraballo, Sharon A. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 120–126.

[8]The smoothed estimates were *not* used to determine which alternative a neighbor voted for; doing so resulted in a 10 percentage-point increase in average error rate.

Cardie, Claire. 1993. A case-based approach to knowledge acquisition for domain-specific sentence analysis. In *Proceedings of the 11th National Conference on Artifical Intelligence*, pages 798–803. AAAI Press/MIT Press.

Chen, Stanley F. and Joshua T. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *34th Annual Meeting of the ACL*, pages 310–318.

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.

Cover, Thomas and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Cover, Thomas M. and Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley series in telecommunications. Wiley-Interscience, New York.

Daelemans, Walter, Antal van den Bosch, and Jakub Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning*, 34(1-3):11–42.

Dagan, Ido, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69.

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

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*, volume 278 of *Kluwer International Series in Engineering and Computer Science*. Kluwer, Boston, July.

Grishman, Ralph and John Sterling. 1993. Smoothing of automatically generated selectional constraints. In *Human Language Technology*, pages 254–259, San Francisco, California. Advanced Research Projects Agency, Software and Intelligent Systems Technology Office, Morgan Kaufmann.

Hatzivassiloglou, Vasileios. 1996. Do we need linguistics when we have statistics? a comparative analysis of the contributions of linguistic cues to a statistical word grouping system. In Judith Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press, chapter 4, pages 67–94.

Hofmann, Thomas and Jan Puzicha. 1998. Mixture models for cooccurrence data. In *International Conference on Pattern Recognition*. Longer version available as MIT A.I. Memo No. 1625 or C.B.C.L. Memo No. 159.

Jelinek, Frederick. 1997. *Statistical Methods for Speech Recognition*. Language Speech, and Communication series. MIT Press.

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.

Kneser, Reinhard and Hermann Ney. 1993. Improved clustering techniques for class-based statistical language modelling. In *European Conference on Speech Communications and Technology*, pages 973–976, Berlin, Germany.

Lapata, Maria. 2000. The automatic interpretation of nominalizations. In *Proceedings of the AAAI*.

Lee, Lillian. 1999. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Lee, Lillian and Fernando Pereira. 1999. Distributional similarity models: Clustering vs. nearest neighbors. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 33–40.

Liebetrau, Albert M. 1983. *Measures of Association*. Number 07-032 in Sage University Paper series on Quantitative Applications in the Social Sciences. Sage Publications, Beverly Hills and London.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL '98*, pages 768–773.

Lin, Jianhua. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, January.

McCarthy, Diana. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 256–263.

Ng, Hwee Tou and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *34th Annual Meeting of the ACL*, pages 40–47.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *31st Annual Meeting of the ACL*, pages 183–190. ACL, Association for Computational Linguistics, Somerset, NJ. Distributed by Morgan Kaufmann, San Francisco, CA.

Rooth, Mats, Stefan Riezler, Detlaf Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 104–111.

Rosenfeld, Ronald. 1996. A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, 10:187–228.

Saul, Lawrence and Fernando C. N. Pereira. 1997. Aggregate and mixed-order Markov models for statistical language processing. In Claire Cardie and Ralph Weischedel, editors, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 81–89, Providence, RI, August.

Stanfill, Craig. 1987. Memory-based reasoning applied to English pronunciation. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 577–581.

Stanfill, Craig and David L. Waltz. 1986. Toward memory-based reasoning. *CACM*, 29(12):1213–1228, December.

Sugawara, K., M. Nishimura, K. Toshioka, M. Okochi, and T. Kaneko. 1985. Isolated word recognition using hidden Markov models. In *Proceedings of ICASSP*, pages 1–4, Tampa, Florida. IEEE.