On the Robustness of Entropy-Based Similarity Measures in Evaluation of Subcategorization Acquisition Systems

Anna Korhonen
University of Cambridge
Computer Laboratory
15 JJ Thomson Avenue
Cambridge CB3 0FD, UK
Anna.Korhonen@cl.cam.ac.uk

Yuval Krymolowski
Bar-Ilan University
Department of Computer Science
Ramat Gan 52900, Israel
yuvalk@cs.biu.ac.il

Abstract

Some statistical learning systems are evaluated using measures of distributional similarity. To deal with the problem of zero events in the distributions under comparison, smoothing is frequently performed before similarity measures are applied. Smoothing alters the information in the original distribution, and may add noise to the results. Here, we investigate the sensitivity of entropy-based similarity measures to noise from uninformative smoothing. Our experiments with two subcategorization acquisition systems show that similarity measures vary in their robustness. While some are led astray by noise from smoothing, others are more resilient.

1 Introduction

Many natural language processing (NLP) tasks involve measuring distributional similarity. Some examples are the estimation of word co-occurrence probabilities (Dagan et al., 1999), automatic construction of thesauri (Lin, 1998), automatic detection of diathesis alternations (McCarthy, 2000), disambiguation of nominalizations (Lapata, 2002), and evaluation of statistical NLP learners (Carroll and Rooth, 1998; Korhonen, 2002b).

Various similarity measures have been proposed and used for NLP purposes, including the Kullback-Leibler distance (Cover and Thomas, 1991), cross entropy (Cover and Thomas, 1991), the Jensen-Shannon divergence (Lin, 1991), the skew divergence (Lee, 1999), cosine (Frakes and Baeza-Yates, 1992), Jaccard's coefficient, L_1 norm, and the confusion probability.

In this paper, we discuss the use of similarity measures in evaluation of the type of statistical language learners which deliver as an output a distribution of events. A typical example is an automatic subcategorization acquisition system (e.g. (Briscoe and Carroll, 1997; Carroll and Rooth, 1998; Korhonen, 2002b)), which learns, from corpus data, a distribution of subcategorization frames (SCFs) specific to a certain predicate ($p(scf_i|predicate_j)$). These learners are frequently evaluated using the standard precision, recall and accuracy measures (Manning and Schütze, 1999). However, similarity measures provide an important means to evaluate the actual acquired frequencies.

In similarity-based evaluation, a learned distribution is compared with a gold standard distribution in order to determine how closely the two correlate. Such evaluation is complicated by cases where the distributions under comparison have different *supports*, i.e. regions of positive probability. Due to the sparse data problem, zero events typically make up a substantial portion of joint data.

To allow the comparison of all events, smoothing is frequently performed before similarity measures are applied. Smoothing tackles the problem by assigning non-zero values to zero events. It is usually done in an *uninformative* way (by assigning a uniform prior probabability to events; e.g. (Laplace, 1814; Witten and Bell, 1991)), rather than in an *informative* way (by assigning an informative prior probability e.g. by backing-off (Katz, 1987)), as it is desirable – from the evaluation point of view – to preserve as much of the original, learned distribution as possible[1].

Although uninformative smoothing makes the simplest possible assumption regarding the probability of unseen events, it involves also altering the information included in the original

[1]See Manning and Schütze (1999) for both informative and uninformative smoothing methods.

distribution. It can add 'noise' to the original data. This noise may, in turn, affect distributional similarity and potentially obscure similarity-based evaluation.

Here, we investigate the sensitivity of widely-used entropy-based similarity measures to the noise from uninformative smoothing. We do this in the context of evaluation, by using the measures to evaluate the accuracy of verbal SCF distributions learned automatically from corpus data. By controlling the number of SCFs smoothed and examining the effect on distributional similarity, we observe differences in the robustness of various measures. Our results show that some entropy-based similarity measures are led astray by noise from smoothing, while others are more resilient and thus better suited for evaluation purposes.

Section 2 introduces the subcategorization learners employed. The similarity measures are described in section 3 and the smoothing methods in section 4. Section 5 reports our experiments. We discuss our findings in section 6 and present our conclusions in section 7.

2 Subcategorization Learners

We used two subcategorization learners proposed by Korhonen (2002b) to obtain the SCF distributions employed in our experiments. The learners are variations of the subcategorization acquisition system of Briscoe and Carroll (1997). The system uses a shallow parser to obtain the subcategorization information from corpus data. It distinguishes 163 verbal SCFs and returns relative frequencies for each SCF found for a given verb. The resulting putative SCF distributions are processed by the two learners as follows:

Learner 1: The SCFs in putative distributions are simply ranked in the order of the probability of their occurrence with the verb. The probabilities are estimated using a maximum likelihood estimate (MLE) from the observed relative frequencies.

Learner 2: The SCF distributions obtained using Learner 1 are smoothed using linear interpolation (Chen and Goodman, 1996). The informative back-off distribution employed in smoothing is based on the semantic class of the verb in question ($p(scf_i|semantic\,class_j)$), chosen according to the verb's most frequent sense in WordNet (Miller, 1990). For instance, the predominant sense of the verb *fly* in WordNet belongs to the semantic class of "Motion" verbs - hence, the distribution of "Motion" verbs is employed as a back-off distribution in smoothing.

The semantic classes are based on Levin classes (Levin, 1993) and the back-off distribution for each class is obtained by merging SCF distributions of a few verbs in the same class. The parameters used in smoothing are obtained by optimising SCF acquisition performance on held-out training data so that most of the smoothed probability is determined by the MLE from the subcategorization acquisition system[2].

Learner 2 tends to perform better than learner 1, since it involves using a priori knowledge about generalizations of verb semantics to guide subcategorization acquisition. It corrects the putative SCF distribution and deals better with sparse data.

Korhonen (2002a) used an empirically determined threshold on the probability estimates to filter noisy SCFs out, and then evaluated the two learners on a test set of 91 verbs using precision and recall -based evaluation. The gold standard employed in the evaluation was obtained by manually analysing an average of 300 occurrences of each test verb in corpus data.

Learner 1 yielded 82% type precision (the percentage of SCF types that the method proposes which are correct) and 49% type recall (the percentage of SCF types in the gold standard that the method proposes), while type precision was 81% and type recall 73% for learner 2. F-measure[3] was thus 61 for learner 1 and 76 for learner 2.

While this evaluation allows us to evaluate the set of acquired SCF types, similarity-based evaluation is needed to evaluate the frequen-

[2] It is important to note that we regard the informative smoothing employed by learner 2 as part of the method for SCF acquisition. When we discuss the effect of smoothing on similarity measures, we essentially mean the uninformative smoothing performed on the *output* of learner 2 for evaluation purposes, not the informative smoothing performed for SCF acquisition.

[3] $F = \frac{2 \cdot precision \cdot recall}{precision + recall}$

cies associated with the types. For example, a learner may correctly acquire a sentential complement frame for *believe* (*I believe that our theory is correct*) but incorrectly assign this frequently occurring frame a very small probability.

3 Similarity Measures

We used the following set of similarity measures to evaluate the accuracy of a learned SCF distribution $q = \{q_i\}$ with respect to a gold standard distribution $p = \{p_i\}$. q_i and p_i denote the probability of scf_i in the two distributions, respectively.

1. **IS:** The intersection measure (Lin, 1998)

$$\textsc{is}(p,q) = \frac{2 \cdot |\mathrm{com}(p,q)|}{|\mathrm{supp}(p)| + |\mathrm{supp}(q)|},$$

where supp(p) and supp(q) are the sets of SCFs with non-zero probability in p and q, and com(p,q) is the intersection of these two sets.

2. **RC:** The Spearman rank correlation coefficient (Spearman, 1904). It involves (i) calculating the ranks r^p and r^q for each of the SCF variables separately, using averaged ranks for tied values, and (ii) finding RC by calculating the Pearson correlation coefficient for the ranks:

$$\textsc{rc}(p,q) = \mathrm{corr}(r^p, r^q).$$

RC lies in the range [-1, 1], with values near 0 denoting a low degree of association and values near -1 and 1 denoting strong association.

3. **CE:** Cross entropy – a measure of the information needed to describe a true distribution p using a model distribution q:

$$\textsc{ce}(p,q) = \sum_i -p_i \log(q_i).$$

CE is minimal when p and q are identical. In this case $\textsc{ce}(p,q) = H(p)$ is the Shannon entropy of p.

4. **KL:** Kullback-Leibler distance – a measure of the additional information needed to describe p using q:

$$D(p\|q) = \textsc{ce}(p,q) - H(p) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right).$$

KL is always ≥ 0 and $= 0$ only when $p \equiv q$.

5. **JS:** The Jensen-Shannon divergence – a measure which relies on the assumption that if p and q are similar, they are close to their average.

$$\textsc{js}(p,q) = \frac{1}{2}\left[D\left(p\middle\|\frac{p+q}{2}\right) + D\left(q\middle\|\frac{p+q}{2}\right)\right].$$

6. **SD:** The skew divergence. It smooths q by mixing it with p:

$$\textsc{sd}(p,q) = D(p\|\alpha \cdot q + (1-\alpha) \cdot p).$$

$\textsc{sd}(p,q)$ approximates KL as $\alpha \to 1$. Lee (1999) reports the best results with $\alpha = 0.99$. We adopted the same value.

All these other measures, except IS and RC (which are included in our experiment primarily for comparison), are entropy-based measures of distributional similarity. CE and KL are asymmetric and, unlike JS and SD, undefined if there exists a SCF for which $p_i > 0$ but $q_i = 0$.

4 Smoothing

Two different uninformative smoothing methods were selected for investigation: the add-one and Witten-Bell (Witten and Bell, 1991) methods. They both work by distributing a certain probability mass among unseen events and discounting the observed distribution accordingly, but differ in the way they estimate the discount.

4.1 Add-One

Add-one smoothing involves assigning a uniform prior probability to all events so that $q_i > 0$ for all i. Let $c(scf_i)$ be the frequency of a SCF (given a verb) in q, N the total number of SCF tokens for this verb in q, and n_{scf} the total number of SCF types[4] considered. The estimated probability of the SCF is:

[4]While 'types' are the set of SCFs assumed e.g. in a dictionary, 'tokens' are the individual occurrences of SCFs e.g. in corpus data.

$$P(scf_i) = \frac{c(scf_i) + 1}{N + n_{\text{scf}}}.$$

4.2 Witten-Bell

Witten and Bell (1991) present a set of smoothing methods which involve estimating the discount from the observed distribution. We adopt their method "C", the so-called Witten-Bell method. It considers each unseen SCF type as an event in addition to the N seen SCF tokens. Accordingly, the probability of an unseen SCF is estimated by

$$p_{\text{wb}} = \frac{n_{\text{scf}}}{N + n_{\text{scf}}}.$$

5 Experiment

5.1 Test Data and Methods

We selected 17 test verbs for our experiments. Sentences containing an occurrence of one of these verbs were first extracted from 20 million words of the British National Corpus (BNC) (Leech, 1992), an average of 1000 citations of each, and then processed using the two subcategorization learners. This yielded two SCF distributions per test verb.

The similarity measures introduced in the previous section were then applied to evaluate the accuracy of these acquired distributions (q) against gold standard distributions (p). The latter were obtained by manually analysing an average of 300 occurrences of each test verb in the BNC.

Prior to calculating the similarity, we smoothed the distributions under comparison using the two uninformative methods introduced in section 4. To investigate in detail the effect of smoothing, we varied the number of SCFs considered, i.e. the number of SCFs which enter into the similarity measurement. This allowed us to control the number of SCFs which require smoothing. Three options were explored which involved considering only those SCFs, for which

Option 1: $p_i > 0$ and $q_i > 0$,

Option 2: $p_i > 0$,

Option 3: either $p_i > 0$ or $q_i > 0$.

Option 1 involves considering the smallest number of SCFs – only those common for p_i and q_i – and never requires smoothing. Options 2 and 3 involve smoothing the gold standard SCFs absent in q_i. Option 3 involves, in addition, smoothing the SCFs which occur in q_i but are absent in p_i.

Option 2 is perhaps the most conventional one when similarity measures are used in evaluation. It involves evaluating the gold standard events only. Option 3 takes into account the false positive events in q_i as well. It involves assigning them a very small probability in p_i, accounting for the fact that these events are extremely unlikely to appear in the gold standard.

5.2 Results

	Add-One		W-B	
	L1	L2	L1	L2
IS				
1.	1.00	1.00	1.00	1.00
2.	0.87	0.96	0.87	0.96
3.	0.36	0.49	0.36	0.49
RC				
1.	0.51	0.78	0.51	0.78
2.	0.48	0.72	0.49	0.72
3.	0.31	0.68	0.31	0.68
CE				
1.	2.01	1.81	2.01	1.81
2.	3.24	2.23	2.38	1.99
3.	3.42	2.30	2.57	2.21
KL				
1.	0.56	0.21	0.56	0.21
2.	1.56	0.56	0.70	0.31
3.	1.74	0.62	0.64	0.31
JS				
1.	0.11	0.05	0.11	0.05
2.	0.14	0.06	0.13	0.06
3.	0.20	0.08	0.13	0.07
SD				
1.	0.53	0.20	0.53	0.20
2.	0.90	0.33	0.66	0.29
3.	1.08	0.40	0.61	0.30

Table 1: Results for the two learners, with the different similarity measures and options 1-3. Results are reported separately for add-one and Witten-Bell smoothing methods.

Table 1 shows the average results for the 17 verbs with each similarity measure and smoothing option (the options 1-3) combination. According to all results reported, learner 2 (L2) is more accurate than learner 1 (L1). The results with IS with option 2 indicate that learner 2 is good in detecting SCFs, finding 92% of the gold standard SCFs, while learner 1 finds only 77% of the SCFs[5]. Thus the number of SCFs smoothed is always higher for learner 1 than for learner 2, and therefore we expect the effect of smoothing to be always stronger for learner 1.

With add-one smoothing, all the entropy-based similarity measures (CE, KL, JS, and SD) show worse results when the number of SCFs smoothed increases. Thus option 1 yields the best results and option 3 the worst. The effect of smoothing is indeed always stronger for learner 1 than learner 2. Interestingly, it also varies largely from one entropy-based measure to another.

KL and CE prove the measures most sensitive to add-one smoothing. When we consider the results for learner 1 and observe the decline in results from option 1 to option 3, KL worsens by a factor of 3.1. CE proves nearly as sensitive. SD worsens by a factor of 2 and JS by a factor of 1.8. From the entropy-based measures, JS is thus the one most resistant to the effect of add-one smoothing. It shows results consistent with those obtained using RC. Similar observations regarding the robustness of the measures can be made with the more accurate learner 2.

With Witten-Bell method, option 1 yields the best results as well. The results for option 2 are not, however, considerably better than those for option 3. In fact with learner 1, one measure – KL – shows the best results with option 3.

This happens because Witten-Bell smoothing tends to assign a higher probability to unseen SCFs than add-one smoothing. Whenever an unseen SCF has a high probability in the distribution where it is seen, Witten-Bell method makes a better guess. This affects the distributional similarity favourably. The converse happens whenever an unseen SCF has a low probability. Although most unseen SCFs have a low probability, the few high probability SCFs have more effect, as entropy-based measures give them more weight. Hence the small differences in results between options 1-3.

Whether or not this indicates that Witten-Bell method is more suitable for our task than add-one method is difficult to judge. However, the results do show that KL is the measure most effected by Witten-Bell smoothing.

Overall, the more sensitive measures (KL and CE) behave similarly with the more robust measures (SD and JS) when option 1 is used. This seems to suggest that from all options, option 1 is the most suitable for these measures when a learned distribution is sparse and a high number of SCFs require smoothing (this is frequently the case with learner 1). Although considering only the SCFs for which $p_i > 0$ and $q_i > 0$ means ignoring a number of events in p_i, it avoids the noise from smoothing and yields more reliable results.

Whether or not the latter unconventional use of similarity measures is generally applicable is a matter which requires further investigation. We conducted our experiments by comparing distributions that share, on average, around half of their SCFs. Further research is required to investigate the effect of smoothing on distributions that share fewer of their SCFs.

6 Discussion

In our experiment, entropy-based similarity measures which require $q_i > 0$ (KL and CE) proved more sensitive to the noise from uninformative smoothing, while those which do not require $q_i > 0$ (SD and JS) proved more robust.

Interestingly, Carroll and Rooth (1998) made similar observations when using CE in evaluation of their subcategorization learner. They noted that uninformative smoothing (using the Poisson model (Witten and Bell, 1991)) introduces a high penalty on CE. They did not investigate other similarity measures or smoothing methods.

Lee (1999) who compared the performance of a variety of similarity measures on a pair co-occurrence task, also reported best results with measures which concentrate effort on events for which both probability estimates are non-zero. She only considered two entropy-based measures – JS and SD – from which SD proved more accurate. Lee (2001) reported better results with SD than KL, even when highly sophisticated (informative) methods were used for smoothing.

We restricted our investigation to entropy-based similarity measures. In the future, it would be interesting to examine the effect of smoothing on commonly used non-entropy

[5] With option 2, the fraction of detected frames is $\frac{\text{IS}}{2-\text{IS}}$.

based similarity measures as well. For example, L_1 (Manhattan) norm and Jaccard's coefficient seem promising candidates on the basis of Lee's (1999) evaluation.

In the future, we also plan to investigate other uninformative smoothing methods (e.g. Poisson and Good-Turing (Good, 1953)), and carry out experiments with a wider range of learners with varying degree of accuracy. Our experiments with the two learners substantially different in their accuracy were inadequate to establish whether the noise from smoothing can actually obscure inter-system comparison.

In this paper, we have examined the use of similarity measures in evaluation of (particular type of) language learners. The results reported suggest that it is best to either evaluate these language learners using similarity measures known to be resistant to noise from smoothing, or possibly employ measures such as KL and CE in an unconventional way, without smoothing: by considering only the nonzero gold standard events in learned distributions.

The SCF distributions we experimented with are typical zipf like ones, which we encounter frequently in natural language. Also, the topic we have investigated is not specific to evaluation: similarity measures are frequently applied to smoothed estimates in other domains as well. Therefore, our observations are likely to be of an interest to the range of NLP tasks which – one way or the other – involve measuring distributional similarity.

7 Conclusion

Uninformative smoothing is frequently performed before the accuracy of automatically acquired SCF distributions is evaluated using measures of distributional similarity. Smoothing allows the comparison of unseen events but adds noise to the original data. In this paper, we investigated the effect of uninformative smoothing on entropy-based similarity measures. We applied these measures to evaluate the accuracy of two subcategorization learners, and studied the effect of smoothing by controlling the number of SCFs considered. We observed variation in the robustness of the different measures. Some measures proved highly sensitive to the noise from smoothing, while others proved more robust and thus preferable for evaluation purposes.

8 Acknowledgements

We thank Ido Dagan and Diana McCarthy for useful comments on this paper. This work was partly supported by UK EPSRC project GR/N36462/93: 'Robust Accurate Statistical Parsing (RASP)'.

References

E. J. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington DC.

Glenn Carroll and Mats Rooth. 1998. Valence induction with a head-lexicalized PCFG. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, Granada, Spain.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the ACL-96*, pages 310–318, Santa Cruz, CA.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of information theory.* Wiley, New York.

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

William B. Frakes and Ricardo Baeza-Yates. 1992. *Information Retrieval: Data Structures and Algorithms.* Prentice Hall, Englewood Cliffs, NJ.

I. J. Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:16–264.

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

Anna Korhonen. 2002a. Semantically motivated subcategorization acquisition. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition.* To appear.

Anna Korhonen. 2002b. *Subcategorization Acquisition.* Ph.D. thesis, University of Cambridge, UK.

Maria Lapata. 2002. The disambiguation of nominalisations. *Computational Linguistics*, 28(3).

P. S. Laplace. 1814. *Essai philosophique sur les probabilites*. Mme. Ve. Courcier.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 25–32.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics 2001*, pages 65–72.

Geoff Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Beth Levin. 1993. *English Verb Classes and Alternations*. Chicago University Press, Chicago.

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

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

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Diana McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the NAACL*, pages 162–169, Seattle, WA.

George A. Miller. 1990. WordNet: An online lexical database. *International Journal of Lexicography*, 3(4):235–312.

C. Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

I. H. Witten and T. C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094.