Estimating and Exploiting the Entropy of Sense Distributions

Peng Jin
Institute of Computational Linguistics
Peking University
Beijing China
jandp@pku.edu.cn

Diana McCarthy, Rob Koeling and **John Carroll**
University of Sussex
Falmer, East Sussex
BN1 9QJ, UK
{dianam,robk,johnca}@sussex.ac.uk

Abstract

Word sense distributions are usually skewed. Predicting the extent of the skew can help a word sense disambiguation (WSD) system determine whether to consider evidence from the local context or apply the simple yet effective heuristic of using the first (most frequent) sense. In this paper, we propose a method to estimate the entropy of a sense distribution to boost the precision of a first sense heuristic by restricting its application to words with lower entropy. We show on two standard datasets that automatic prediction of entropy can increase the performance of an automatic first sense heuristic.

1 Introduction

Word sense distributions are typically skewed and WSD systems do best when they exploit this tendency. This is usually done by estimating the most frequent sense (MFS) for each word from a training corpus and using that sense as a back-off strategy for a word when there is no convincing evidence from the context. This is known as the MFS heuristic [1] and is very powerful since sense distributions are usually skewed. The heuristic becomes particularly hard to beat for words with highly skewed sense distributions (Yarowsky and Florian, 2002). Although the MFS can be estimated from tagged corpora, there are always cases where there is insufficient data, or where the data is inappropriate, for example because it comes from a very different domain. This has motivated some recent work attempting to estimate the distributions automatically (McCarthy et al., 2004; Lapata and Keller, 2007). This paper examines the case for determining the skew of a word sense distribution by estimating entropy and then using this to increase the precision of an unsupervised first sense heuristic by restricting application to those words where the system can automatically detect that it has the most chance. We use a method based on that proposed by McCarthy et al. (2004) as this approach does not require hand-labelled corpora. The method could easily be adapted to other methods for predicing predominant sense.

[1] It is also referred to as the first sense heuristic in the WSD literature and in this paper.

2 Method

Given a listing of senses from an inventory, the method proposed by McCarthy et al. (2004) provides a prevalence ranking score to produce a MFS heuristic. We make a slight modification to McCarthy et al.'s prevalence score and use it to estimate the probability distribution over the senses of a word. We use the same resources as McCarthy et al. (2004): a distributional similarity thesaurus and a WordNet semantic similarity measure. The thesaurus was produced using the metric described by Lin (1998) with input from the grammatical relation data extracted using the 90 million words of written English from the British National Corpus (BNC) (Leech, 1992) using the RASP parser (Briscoe and Carroll, 2002). The thesaurus consists of entries for each word (w) with the top 50 "nearest neighbours" to w, where the neighbours are words ranked by the distributional similarity that

Proceedings of NAACL HLT 2009: Short Papers, pages 233–236,
Boulder, Colorado, June 2009.

they share with w. The WordNet similarity score is obtained with the **jcn** measure (Jiang and Conrath, 1997) using the WordNet Similarity Package 0.05 (Patwardhan and Pedersen, 2003) and WordNet version 1.6. The **jcn** measure needs word frequency information, which we obtained from the BNC.

2.1 Estimates of Predominance, Probability and Entropy

Following McCarthy et al. (2004), we calculate prevalence of each sense of the word (w) using a weighted sum of the distributional similarity scores of the top 50 neighbours of w. The sense of w that has the highest value is the automatically detected MFS (predominant sense). The weights are determined by the WordNet similarity between the sense in question and the neighbour. We make a modification to the original method by multiplying the weight by the inverse rank of the neighbour from the list of 50 neighbours. This modification magnifies the contribution to each sense depending on the rank of the neighbour while still allowing a neighbour to contribute to all senses that it relates too. We verified the effect of this change compared to the original ranking score by measuring cross-entropy. [2]

Let $N_w = n_1, n_2 \ldots n_k$ denote the ordered set of the top $k = 50$ neighbours of w according to the distributional similarity thesaurus, $senses(w)$ is the set of senses of w and $dss(w, n_j)$ is the distributional similarity score of a word w and its j^{th} neighbour. Let ws_i be a sense of w then $wnss(ws_i, n_j)$ is the maximum WordNet similarity score between ws_i and the WordNet sense of the neighbour (n_j) that maximises this score. The prevalence score is calculated as follows with $\frac{1}{rank_{n_j}}$ being our modification to McCarthy et al.

$$Prevalence\ Score(ws_i) = \sum_{n_j \in N_w} dss(w, n_j) \times \frac{wnss(ws_i, n_j)}{\sum_{ws_{i'} \in senses(w)} wnss(ws_{i'}, n_j)} \times \frac{1}{rank_{n_j}} \quad (1)$$

To turn this score into a probability estimate we sum the scores over all senses of a word and the probability for a sense is the original score divided by this sum:

$$\hat{p}(ws_i) = \frac{prevalence\ score(ws_i)}{\sum_{ws_j \in w} prevalence\ score(ws_j)} \quad (2)$$

To smooth the data, we evenly distribute 1/10 of the smallest prevalence score to all senses with a undefined prevalence score values. Entropy is measured as:

$$H(senses(w)) = - \sum_{ws_i \in senses(w)} p(ws_i) log(p(ws_i))$$

using our estimate ($\hat{p}$) for the probability distribution p over the senses of w.

3 Experiments

We conducted two experiments to evaluate the benefit of using our estimate of entropy to restrict application of the MFS heuristic. The two experiments are conducted on the polysemous nouns in SemCor and the nouns in the SENSEVAL-2 English all words task (we will refer to this as SE2-EAW).

3.1 SemCor

For this experiment we used all the polysemous nouns in Semcor 1.6 (excluding multiwords and proper nouns). We depart slightly from (McCarthy et al., 2004) in including all polysemous nouns whereas they limited the experiment to those with a frequency in SemCor of 3 or more and where there is one sense with a higher frequency than the others. Table 1 shows the precision of finding the predominant sense using equation 1 with respect to different entropy thresholds. At each threshold, the MFS in Semcor provides the upper-bound (UB). The random baseline (RBL) is computed by selecting one of the senses of the target word randomly as the predominant sense. As we hypothesized, precision is higher when the entropy of the sense distribution is lower, which is an encouraging result given that the entropy is automatically estimated. The performance of the random baseline is higher at lower entropy which shows that the task is easier and involves a lower degree of polysemy of the target words. However, the gains over the random baseline are greater at lower entropy levels indicating that the merits of detecting the skew of the distribution cannot all be due to lower polysemy levels.

[2] Our modified version of the score gave a lower cross-entropy with SemCor compared to that in McCarthy et al. The result was highly significant with $p < 0.01$ on the t-test.

H ($\leq$)	precision eq 1	RBL	UB	# tokens
0.5	-	-	-	0
0.9	80.3	50.0	84.8	466
0.95	85.1	50.0	90.9	1360
1	68.5	50.0	87.4	9874
1.5	67.6	42.6	86.9	11287
2	58.0	36.7	79.5	25997
2.5	55.7	34.4	77.6	31599
3.0	50.2	30.6	73.4	41401
4.0	47.6	28.5	70.8	46987
5.0 (all)	47.3	27.3	70.5	47539

Table 1: First sense heuristic on SemCor

Freq $\leq$	P	#tokens
1	45.9	1132
5	50.1	5765
10	50.7	10736
100	49.4	39543
1000(all)	47.3	47539
#senses $\leq$	P	#tokens
2	67.2	10736
5	55.4	31181
8	50.1	41393
12	47.8	46041
30(all)	47.3	47539

Table 2: Precision (P) of equation 1 on SemCor with respect to frequency and polysemy

We also conducted a frequency and polysemy analysis shown in Table 2 to demonstrate that the increase in precision is not all due to frequency or polysemy. This is important, since both frequency and polysemy level (assuming a predefined sense inventory) could be obtained without the need for automatic estimation. As we can see, while precision is higher for lower polysemy, the automatic estimate of entropy can provide a greater increase in precision than polysemy, and frequency does not seem to be strongly correlated with precision.

3.2 SENSEVAL-2 English All Words Dataset

The SE2-EAW task provides a hand-tagged test suite of 5,000 words of running text from three articles from the Penn Treebank II (Palmer et al., 2001). Again, we examine whether precision of the MFS heuristic can be increased by restricting application depending on entropy. We use the same resources as for the SemCor experiment. [3] Table 3 gives the results. The most frequent sense (MFS) from SE2-EAW itself provides the upper-bound (UB). We also compare performance with the Semcor MFS (SC). Performance is close to the Semcor MFS while not relying on any manual tagging. As before, precision increases significantly for words with low estimated entropy, and the gains over the random baseline are higher compared to the gains including all words.

H ($\leq$)	precision eq 1	RBL	SC	UB	# tokens
0.5	-	-	-	-	0
0.9	1	50.0	1	1	7
0.95	94.7	50.0	94.7	1	19
1	69.6	50.0	81.3	94.6	112
1.5	68.0	49.0	81.3	93.8	128
2	69.6	34.7	68.2	87.7	421
2.5	65.0	33.0	65.0	86.5	488
3.0	56.6	27.5	60.8	80.1	687
4.0	52.6	25.6	58.8	79.2	766
5.0 (all)	51.5	25.6	58.5	79.3	769

Table 3: First sense heuristic on SE2-EAW

4 Related Work

There is promising related work on determining the predominant sense for a MFS heuristic (Lapata and Keller, 2007; Mohammad and Hirst, 2006) but our work is the first to use the ranking score to estimate entropy and apply it to determine the confidence in the MFS heuristic. It is likely that these methods would also have increased precision if the ranking scores were used to estimate entropy. We leave such investigations for further work.

Chan and Ng (2005) estimate word sense distributions and demonstrate that sense distribution estimation improves a supervised WSD classifier. They use three sense distribution methods, including that of McCarthy et al. (2004). While the other two methods outperform the McCarthy et al. method,

[3] We also used a tool for mapping from WordNet 1.7 to WordNet 1.6 (Daudé et al., 2000) to map the SE2-EAW noun data (originally distributed with 1.7 sense numbers) to 1.6 sense numbers.

they rely on parallel training data and are not applicable on 9.6% of the test data for which there are no training examples. Our method does not require parallel training data.

Agirre and Martínez (2004) show that sense distribution estimation is very important for both supervised and unsupervised WSD. They acquire tagged examples on a large scale by querying Google with monosemous synonyms of the word senses in question. They show that the method of McCarthy et al. (2004) can be used to produce a better sampling technique than relying on the bias from web data or randomly selecting the same number of examples for each sense. Our work similarly shows that the automatic MFS is an unsupervised alternative to SemCor but our work does not focus on sampling but on an estimation of confidence in an automatic MFS heuristic.

5 Conclusions

We demonstrate that our variation of the McCarthy et al. (2004) method for finding a MFS heuristic can be used for estimating the entropy of a sense distribution which can be exploited to boost precision. Words which are estimated as having lower entropy in general get higher precision. This suggests that automatic estimation of entropy is a good criterion for getting higher precision. This is in agreement with Kilgarriff and Rosenzweig (2000) who demonstrate that entropy is a good measure of the difficulty of WSD tasks, though their measure of entropy was taken from the gold-standard distribution itself.

As future work, we want to compare this approach of estimating entropy with other methods for estimating sense distributions which do not require hand-labelled data or parallel texts. Currently, we disregard local context. We wish to couple the confidence in the MFS with contextual evidence and investigate application on coarse-grained datasets.

Acknowledgements

This work was funded by the China Scholarship Council, the National Grant Fundamental Research 973 Program of China: Grant No. 2004CB318102, the UK EPSRC project EP/C537262 'Ranking Word Senses for Disambiguation', and a UK Royal Society Dorothy Hodgkin Fellowship to the second author.

References

E. Agirre and D. Martínez. 2004. Unsupervised wsd based on automatically retrieved examples: The importance of bias. In *Proceedings of EMNLP-2004*, pages 25–32, Barcelona, Spain.

E. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of LREC-2002*, pages 1499–1504, Las Palmas, Canary Islands, Spain.

Y.S. Chan and H.T. Ng. 2005. Word sense disambiguation with distribution estimation. In *Proceedings of IJCAI 2005*, pages 1010–1015, Edinburgh, Scotland.

J. Daudé, L. Padró, and G. Rigau. 2000. Mapping wordnets using structural information. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference on Research in Computational Linguistics*, Taiwan.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for english SENSEVAL. *Computers and the Humanities. Senseval Special Issue*, 34(1–2):15–48.

M. Lapata and F. Keller. 2007. An information retrieval approach to sense ranking. In *Proceedings of NAACL-2007*, pages 348–355, Rochester.

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

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 98*, Montreal, Canada.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of ACL-2004*, pages 280–287, Barcelona, Spain.

S. Mohammad and G. Hirst. 2006. Determining word sense dominance using a thesauru s. In *Proceedings of EACL-2006*, pages 121–128, Trento, Italy.

M. Palmer, C. Fellbaum, S. Cotton, L. Delfs, and H. Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of the* SENSEVAL-2 *workshop*, pages 21–24.

S. Patwardhan and T. Pedersen. 2003. The wordnet::similarity package. http://wn-similarity.sourceforge.net/.

D. Yarowsky and R. Florian. 2002. Evaluating sense disambiguation performance across diverse parameter spaces. *Natural Language Engineering*, 8(4):293–310.