Exploring the Sense Distributions of Homographs

Reinhard Rapp
University of Mainz, FASK
76711 Germersheim, Germany
rrapp@uni-mainz.de

Abstract

This paper quantitatively investigates in how far local context is useful to disambiguate the senses of an ambiguous word. This is done by comparing the co-occurrence frequencies of particular context words. First, one context word representing a certain sense is chosen, and then the co-occurrence frequencies with two other context words, one of the same and one of another sense, are compared. As expected, it turns out that context words belonging to the same sense have considerably higher co-occurrence frequencies than words belonging to different senses. In our study, the sense inventory is taken from the University of South Florida homograph norms, and the co-occurrence counts are based on the British National Corpus.

1 Introduction

Word sense induction and disambiguation is of importance for many tasks in speech and language processing, such as speech recognition, machine translation, natural language understanding, question answering, and information retrieval. As evidenced by several SENSEVAL sense disambiguation competitions (Kilgarriff & Palmer, 2000), statistical methods are dominant in this field. However, none of the published algorithms comes close to human performance in word sense disambiguation, and it is therefore unclear in how far the statistical regularities that are exploited in these algorithms are a solid basis to eventually solve the problem.

Although this is a difficult question, in this study we try to give at least a partial answer. Our starting point is the observation that ambiguous words can usually be disambiguated by their context, and that certain context words can be seen as indicators of certain senses. For example, context words such as *finger* and *arm* are typical of the *hand* meaning of *palm*, whereas *coconut* and *oil* are typical of its *tree* meaning. The essence behind many algorithms for word sense disambiguation is to implicitly or explicitly classify all possible context words into groups relating to one or another sense. This can be done in a supervised (Yarowsky, 1994), a semi-supervised (Yarowsky, 1995) or a fully unsupervised way (Pantel & Lin, 2002).

However, the classification can only work if the statistical clues are clear enough and if there are not too many exceptions. In terms of word co-occurrence statistics, we can say that within the local contexts of an ambiguous word, context words typical of the same sense should have high co-occurrence counts, whereas context words associated with different senses should have co-occurrence counts that are considerably lower. Although the relative success of previous disambiguation systems (e.g. Yarowsky, 1995) suggests that this should be the case, the effect has usually not been quantified as the emphasis was on a task-based evaluation. Also, in most cases the amount of context to be used has not been systematically examined.

2 Methodology

Our starting point is a list of 288 ambiguous words (homographs) where each comes together with two associated words that are typical of one sense and a third associated word that is typical of another sense. Table 1 shows the first ten entries in the list. It has been derived from the *University of South Florida homograph norms* (Nelson et al., 1980) and is based on a combination of native speakers' intuition and the expertise of specialists.

The University of South Florida homograph norms comprise 320 words which were all selected from *Roget's International Thesaurus* (1962). Each word has at least two distinct meanings that were judged as likely to be understood by everyone. As described in detail in Nelson et al. (1980), the compilation of the norms was conducted as follows: 46 subjects wrote down the first word that came to mind for each of the 320 homographs. In the next step, for each homograph semantic categories were chosen to reflect

its meanings. All associative responses given by the subjects were assigned to one of these categories. This was first done by four judges individually, and then, before final categorization, each response was discussed until a consensus was achieved.

The data used in our study (first ten items shown in Table 1) was extracted from these norms by selecting for each homograph the first two words relating to its first meaning and the first word relating to its second meaning. Thereby we had to abandon those homographs where all of the subjects' responses had been assigned to a single category, so that only one category appeared in the homograph norms. This was the case for 32 words, which is the reason that our list comprises only 288 instead of 320 items.

Another resource that we use is the *British National Corpus* (BNC), which is a balanced sample of written and spoken English that comprises about 100 million words (Burnard & Aston, 1998). This corpus was used without special preprocessing, i.e. stop words were not removed and no stemming was conducted. From the corpus we extracted concordances comprising text windows of a certain width (e.g. plus and minus 20 words around the given word) for each of the 288 homographs. For each concordance we computed two counts: The first is the number of concordance lines where the two words associated with sense 1 occur together. The second is the number of concordance lines where the first word associated with sense 1 and the word associated with sense 2 co-occur. The expectation is that the first count should be higher as words associated to the same sense should co-occur more often than words associated to different senses.

homograph	sense 1		sense 2
	first association (*w1*)	second association (*w2*)	first association (*w3*)
arm	leg	hand	war
ball	game	base	dance
bar	drink	beer	crow
bark	dog	loud	tree
base	ball	line	bottom
bass	fish	trout	drum
bat	ball	boy	fly
bay	Tampa	water	hound
bear	animal	woods	weight
beam	wood	ceiling	light

Table 1. First ten of 288 homographs and some associations to their first and second senses.

However, as absolute word frequencies can vary over several orders of magnitude and as this effect could influence our co-occurrence counts in an undesired way, we decided to take this into account by dividing the co-occurrence counts by the concordance frequency of the second words in our pairs. We did not normalize for the frequency of the first word as it is identical for both pairs and therefore represents a constant factor. Note that we normalized for the observed frequency within the concordance and not within the entire corpus.

If we denote the first word associated to sense 1 with *w1*, the second word associated with sense 1 with *w2*, and the word associated with sense 2 with *w3*, the two scores *s1* and *s2* that we compute can be described as follows:

$$s1 = \frac{\text{number of lines where } w1 \text{ and } w2 \text{ co-occur}}{\text{occurrence count of } w2 \text{ within concordance}}$$

$$s2 = \frac{\text{number of lines where } w1 \text{ and } w3 \text{ co-occur}}{\text{occurrence count of } w3 \text{ within concordance}}$$

In cases where the denominator was zero we assigned a score of zero to the whole expression. For all 288 homographs we compared *s1* to *s2*. If it turns out that in the vast majority of cases *s1* is higher than *s2*, then this result would be an indicator that it is promising to use such co-occurrence statistics for the assignment of context words to senses. On the other hand, should this not be the case, the conclusion would be that this approach does not have the potential to work and should be discarded.

As in statistics the results are often not as clear cut as would be desirable, for comparison we conducted another experiment to help us with the interpretation. This time the question was whether our results were caused by properties of the homographs or if we had only measured properties of the context words *w1*, *w2* and *w3*. The idea was to conduct the same experiment again, but this time not based on concordances but on the entire corpus. However, considering the entire corpus would make it necessary to use a different kind of text window for counting the co-occurrences as there would be no given word to center the text window around, which could lead to artefacts and make the comparison problematic. We therefore decided to use concordances again, but this time not the concordances of the homographs (first column in Table 1) but the concordances of all 288 instances of *w1* (second column in Table 1). This way we had exactly

the same window type as in the first experiment, but this time the entire corpus was taken into account as all co-occurrences of *w2* or *w3* with *w1* must necessarily appear within the concordance of *w1*.

We name the scores resulting from this experiment *s3* and *s4*, where *s3* corresponds to *s1* and *s4* corresponds to *s2*, with the only difference being that the concordances of the homographs are replaced by the concordances of the instances of *w1*. Regarding the interpretation of the results, if the ratio between *s3* and *s4* should turn out to be similar to the ratio between *s1* and *s2*, then the influence of the homographs would be marginally or non existent. If there should be a major difference, then this would give evidence that, as desired, a property of the homograph has been measured.

3 Results and discussion

Following the procedure described in the previous section, Table 2 gives some quantitative results. It shows the overall results for the homograph-based concordance and for the *w1*-based concordance for different concordance widths. In each case not only the number of cases is given where the results correspond to expectations (*s1* > *s2* and *s3* > *s4*), but also the number of cases where the outcome is undecided (*s1* = *s2* and *s3* = *s4*). Although this adds some redundancy, for convenience also the number of cases with an unexpected outcome is listed. All three numbers sum up to 288 which is the total number of homographs considered.

If we look at the left half of Table 2 which shows the results for the concordances based on the homographs, we can see that the number of correct cases steadily increases with increasing width of the concordance until a width of ±300 is reached. At the same time, the number of undecided cases rapidly goes down. At a concordance width of ±300, the number of correct cases (201) outnumbers the number of incorrect cases (63) by a factor of 3.2. Note that the increase of incorrect cases is probably mostly an artefact of the sparse-data-problem as the number of undecided cases decreases faster than the number of correct cases increases.

On the right half of Table 2 the results for the concordances based on *w1* are given. Here the number of correct cases starts at a far higher level for small concordance widths, increases up to a concordance width of ±10 where it reaches its maximum, and then decreases slowly. At the concordance width of ±10 the ratio between correct and incorrect cases is 2.6.

How can we now interpret these results? What we can say for sure when we look at the number of undecided cases is that the problem of data sparseness is much more severe if we consider the concordances of the homographs rather than the concordances of *w1*. This outcome can be expected as in the first case we only take a (usually small) fraction of the full corpus into account, whereas the second case is equivalent to considering the full corpus. What we can also say is that the optimal concordance width depends on data sparseness. If data is more sparse, we need a wider concordance width to obtain best results.

concordance width	concordance of homograph			concordance of *w1*		
	s1 > *s2* correct	*s1* = *s2* undecided	*s1* < *s2* incorrect	*s3* > *s4* correct	*s3* = *s4* undecided	*s3* < *s4* incorrect
±1	1	287	0	107	135	46
±2	15	273	0	158	69	61
±3	32	249	7	179	40	69
±5	54	222	12	194	21	73
±10	81	181	26	199	13	76
±20	126	127	35	196	7	85
±30	129	105	44	192	5	91
±50	165	69	54	192	2	94
±100	182	44	62	185	1	102
±200	198	29	61	177	1	110
±300	201	24	63	177	1	110
±500	199	19	70	171	1	116

Table 2. Results for homograph-based concordance (left) and for *w1*-based concordance (right).

In case of the full corpus the optimal width is around ±10 which is similar to average sentence length. Larger windows seem to reduce saliency and therefore affect the results adversely. In comparison, if we look at the concordances of the homographs, the negative effect on saliency with increasing concordance width seems to be more than outweighed by the decrease in sparseness, as the results at a very large width of ±300 are better than the best results for the full corpus. However, if we used a much larger corpus than the BNC, it can be expected that best results would be achieved at a smaller width, and that these are likely to be better than the ones achieved using the BNC.

4 Conclusions and future work

Our experiments showed that associations belonging to the same sense of a homograph have far higher co-occurrence counts than associations belonging to different senses. This is especially true when we look at the concordances of the homographs, but – to a somewhat lesser extend – also when we look at the full corpus. The discrepancy between the two approaches can probably be enlarged by increasing the size of the corpus. However, further investigations are necessary to verify this claim.

With the approach based on the concordances of the homographs best results were achieved with concordance widths that are about an order of magnitude larger than average sentence length. However, human performance shows that the context within a sentence usually suffices to disambiguate a word. A much larger corpus could possibly solve this problem as it should allow to reduce concordance width without loosing accuracy. However, since human language acquisition seems to be based on the reception of only in the order of 100 million words (Landauer & Dumais, 1997, p. 222), and because the BNC already is of that size, there also must be another solution to this problem.

Our suggestion is to not look at the co-occurrence frequencies of single word pairs, but at the *average* co-occurrence frequencies between several pairs derived from larger groups of words. Let us illustrate this by coming back to our example in the introduction, where we stated that context words such as *finger* and *arm* are typical of the *hand* meaning of *palm*, whereas *coconut* and *oil* are typical of its *tree* meaning. The sparse-data-problem may possibly prevent our expectation come true, namely that *finger* and *arm* co-occur more often than *finger* and *coconut*. But if we add other words that are typical of the hand meaning, e.g. *hold* or *wrist*, then an incidental lack of observed co-occurrences between a particular pair can be compensated by co-occurrences between other pairs. Since the number of possible pairs increases quadratically with the number of words that are considered, this should have a significant positive effect on the sparse-data-problem, which is to be examined in future work.

Acknowledgments

I would like to thank the three anonymous reviewers for their detailed and helpful comments.

References

Burnard, Lou.; Aston, Guy (1998). *The BNC Handbook: Exploring the British National Corpus with Sara*. Edinburgh University Press.

Kilgarriff, Adam; Palmer, Martha (eds.) (2000). *International Journal of Computers and the Humanities. Special Issue on SENSEVAL*, 34(1-2), 2000.

Landauer, Thomas K.; Dumais, Susan S. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104(2), 211-240.

Nelson, Douglas L.; McEvoy, Cathy L.; Walling, John R.; Wheeler, Joseph W. (1980). The University of South Florida homograph norms. *Behavior Research Methods & Instrumentation* 12(1), 16-37.

Pantel, Patrick; Lin, Dekang (2002). Discovering word senses from text. In: *Proceedings of ACM SIGKDD*, Edmonton, 613-619.

Roget's International Thesaurus (3rd ed., 1962). New York: Crowell.

Yarowsky, David (1994). Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. *Proceedings of the 32nd Meeting of the ACL*, Las Cruces, NM, 88-95.

Yarowsky, David (1995). Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Meeting of the ACL*, Cambridge, MA, 189-196.