Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions

Alpha K. Luk

Microsoft Institute
North Ryde, NSW 2113, Australia
t-alphal@microsoft.com

Department of Computing
Macquarie University
NSW 2109, Australia

Abstract

Corpus-based sense disambiguation methods, like most other statistical NLP approaches, suffer from the problem of data sparseness. In this paper, we describe an approach which overcomes this problem using dictionary definitions. Using the definition-based conceptual co-occurrence data collected from the relatively small Brown corpus, our sense disambiguation system achieves an average accuracy comparable to human performance given the same contextual information.

1 Introduction

Previous corpus-based sense disambiguation methods require substantial amounts of sense-tagged training data (Kelly and Stone, 1975; Black, 1988 and Hearst, 1991) or aligned bilingual corpora (Brown et al., 1991; Dagan, 1991 and Gale et al. 1992). Yarowsky (1992) introduces a thesaurus-based approach to statistical sense disambiguation which works on monolingual corpora without the need for sense-tagged training data. By collecting statistical data of word occurrences in the context of different thesaurus categories from a relatively large corpus (10 million words), the system can identify salient words for each category. Using these salient words, the system is able to disambiguate polysemous words with respect to thesaurus categories.

Statistical approaches like these generally suffer from the problem of data sparseness. To estimate the salience of a word with reasonable accuracy, the system needs the word to have a significant number of occurrences in the corpus. Having large corpora will help but some words are simply too infrequent to make a significant statistical contribution even in a rather large corpus. Moreover, huge corpora are not generally available in all domains and storage and processing of very huge corpora can be problematic in some cases.[1]

In this paper, we describe an approach which attacks the problem of data sparseness in automatic statistical sense disambiguation. Using definitions from LDOCE (Longman Dictionary of Contemporary English; Procter, 1978), co-occurrence data of concepts, rather than words, is collected from a relatively small corpus, the one million word Brown corpus. Since all the definitions in LDOCE are written using words from the 2000 word *controlled vocabulary* (or in our terminology, defining concepts), even our small corpus is found to be capable of providing statistically significant co-occurrence data at the level of the defining concepts. This data is then used in a sense disambiguation system. The system is tested on twelve words previously discussed in the sense disambiguation literature. The results are found to be comparable to human performance given the same contextual information.

2 Statistical Sense Disambiguation Using Dictionary Definitions

It is well known that some words tend to co-occur with some words more often than with others. Similarly, looking at the meaning of the words, one should find that some concepts co-occur more often with some concepts than with others. For example, the concept *crime* is found to co-occur frequently with the concept *punishment*. This kind of conceptual relationship is not always reflected at the lexical level. For instance, in legal reports, the

[1] Statistical data is domain dependent. Data extracted from a corpus of one particular domain is usually not very useful for processing text of another domain.

concept *crime* will usually be expressed by words like *offence* or *felony*, etc., and *punishment* will be expressed by words such as *sentence*, *fine* or *penalty*, etc. The large number of different words of similar meaning is the major cause of the data sparseness problem.

The meaning or underlying concepts of a word are very difficult to capture accurately but dictionary definitions provide a reasonable representation and are readily available.[2] For instance, the LDOCE definitions of both *offence* and *felony* contain the word *crime*, and all of the definitions of *sentence*, *fine* and *penalty* contain the word *punishment*. To disambiguate a polysemous word, a system can select the sense with a dictionary definition containing defining concepts that co-occur most frequently with the defining concepts in the definitions of the other words in the context. In the current experiment, this conceptual co-occurrence data is collected from the Brown corpus.

2.1 Collecting Conceptual Co-occurrence Data

Our system constructs a two-dimensional table which records the frequency of co-occurrence of each pair of defining concepts. The controlled vocabulary provided by Longman is a list of all the words used in the definitions but, in its crude form, it does not suit our purpose. From the controlled vocabulary, we manually constructed a list of 1792 defining concepts. To minimise the size of the table and the processing time, all the closed class words and words which are rarely used in definitions (e.g., the days of the week, the months) are excluded from the list. To strengthen the signals, words which have the same semantic root are combined as one element in the list (e.g., *habit* and *habitual* are combined as {*habit, habitual*}).

The whole LDOCE is pre-processed first. For each entry in LDOCE, we construct its corresponding *conceptual expansion*. The conceptual expansion of an entry whose headword is not a defining concept is a set of *conceptual sets*. Each conceptual set corresponds to a sense in the entry and contains all the defining concepts which occur in the definition of the sense. The entry of the noun *sentence* and its corresponding conceptual expansion are shown in Figure 1. If the headword of an entry is a defining concept DC, the conceptual expansion is given as {{DC}}.

The corpus is pre-segmented into sentences but not pre-processed in any other way (sense-tagged or part-of-speech-tagged). The context of a word is defined to be the current sentence.[3] The system processes the corpus sentence by sentence and collects conceptual co-occurrence data for each defining concept which occurs in the sentence. This allows the whole table to be constructed in a single run through the corpus.

Since the training data is not sense tagged, the data collected will contain noise due to spurious senses of polysemous words. Like the thesaurus-based approach of Yarowsky (1992), our approach relies on the dilution of this noise by their distribution through all the 1792 defining concepts.

Different words in the corpus have different numbers of senses and different senses have definitions of varying lengths. The principle adopted in collecting co-occurrence data is that every pair of content words which co-occur in a sentence should have equal contribution to the conceptual co-occurrence data regardless of the number of definitions (senses) of the words and the lengths of the definitions. In addition, the contribution of a word should be evenly distributed between all the senses of a word and the contribution of a sense should be evenly distributed between all the concepts in a sense. The algorithm for conceptual co-occurrence data collection is shown in Figure 2.

2.2 Using the Conceptual Co-occurrence Data for Sense Disambiguation

To disambiguate a polysemous word W in a context C, which is taken to be the sentence containing W, the system scores each sense S of W, as defined in LDOCE, with respect to C using the following equations.

$$score(S,C) = score(CS,C') - score(CS,GlobalCS) \quad [1]$$

where CS is the corresponding conceptual set of S, C' is the set of conceptual expansions of all content words (which are defined in LDOCE) in C and $GlobalCS$ is the conceptual set containing all the 1792 defining concepts.

---

[2] Manually constructed semantic frames could be more useful computationally but building semantic frames for a huge lexicon is an extremely expensive exercise.

[3] The average sentence length of the Brown corpus is 19.4 words.

Entry in LDOCE	conceptual expansion
1. (an order given by a judge which fixes) a punishment for a criminal found guilty in court	{ {order, judge, punish, crime, criminal, find, guilt, court},
2. a group of words that forms a statement, command, **exclamation**, or question, usu. contains a subject and a verb, and (in writing) begins with a capital letter and ends with one of the marks . ! ?	{group, word, form, statement, command, question, contain, subject, verb, write, begin, capital, letter, end, mark} }

Figure 1. The entry of *sentence* (n.) in LDOCE and its corresponding conceptual expansion

1. Initialise the Conceptual Co-occurrence Data Table (CCDT) with initial value of 0 for each cell.
2. For each sentence S in the corpus, do
   a. Construct S', the set of conceptual expansions of all content words (which are defined in LDOCE) in S.
   b. For each unique pair of conceptual expansions (CE_i, CE_j) in S', do
      For each defining concept DC_{imp} in each conceptual set CS_{im} in CE_i, do
         For each defining concept DC_{jnq} in each conceptual set CS_{jn} in CE_j, do
            increase the values of the cells $CCDT(DC_{imp}, DC_{jnq})$
            and $CCDT(DC_{jnq}, DC_{imp})$ by the product of $w(DC_{imp})$ and $w(DC_{jnq})$
            where $w(DC_{xyz})$ is the weight of DC_{xyz} given by

$$w(DC_{xyz}) = \frac{1}{|CE_x| \bullet |CS_{xy}|}$$

Figure 2. The algorithm for collecting conceptual co-occurrence data

$$score(CS,C') = \sum_{\forall CE' \in C'} score(CS,CE') \Big/ |C'|$$

for any concp. set CS and concp. exp. set C' [2]

$$score(CS,CE') = \max_{CS' \in CE'} score(CS,CS')$$

for any concp. set CS and concp. exp. CE' [3]

$$score(CS,CS') = \sum_{\forall DC' \in CS'} score(CS,DC') \Big/ |CS'|$$

for any concp. sets CS and CS' [4]

$$score(CS,DC') = \sum_{\forall DC \in CS} score(DC,DC') \Big/ |CS|$$

for any concp. set CS and def. concept DC' [5]

$$score(DC,DC') = \max(0, I(DC,DC'))$$

for any def. concepts DC and DC' [6]

$I(DC,DC')$ is the mutual information[4] (Fano, 1961) between the 2 defining concepts DC and DC' given by:

$$I(x,y) \equiv \log_2 \frac{P(x,y)}{P(x) \bullet P(y)}$$

$$\approx \log_2 \frac{f(x,y) \bullet N}{f(x) \bullet f(y)}$$

(using the Maximum Likelihood Estimator).

$f(x,y)$ is looked up directly from the conceptual co-occurrence data table, $f(x)$ and $f(y)$ are looked up from a pre-constructed list of $f(DC)$ values, for each defining concept DC:

$$f(DC) = \sum_{\forall DC'} f(DC,DC')$$

[4] Church and Hanks (1989) use Mutual Information to measure word association norms.

N is taken to be the total number of pairs of words processed, given by

$$\sum_{\forall DC} f(DC) \Big/ 2$$

since for each pair of surface words processed,

$$\sum_{\forall DC} f(DC)$$

is increased by 2.

Our scoring method is based on a probabilistic model at the conceptual level. In a standard model, the logarithm of the probability of occurrence of a conceptual set $\{x_1, x_2, ..., x_m\}$ in the context of the conceptual set $\{y_1, y_2, ..., y_n\}$ is given by

$$\log_2 P(x_1, x_2, ..., x_m | y_1, y_2, ..., y_n)$$

$$\approx \sum_{i=1}^{m} \left(\sum_{j=1}^{n} I(x_i, y_j) + \log_2 P(x_i) \right)$$

assuming that each $P(x_i)$ is independent of each other given $y_1, y_2, ..., y_n$ and each $P(y_j)$ is independent of each other given x_i, for all x_i.[5]

Our scoring method deviates from the standard model in a number of aspects:

1. $\log_2 P(x_i)$, the term of the occurrence probability of each of the defining concepts in the sense, is excluded in our scoring method. Since the training data is not sense-tagged, the occurrence probability is highly unreliable. Moreover, the magnitude of mutual information is decreased due to the noise of the spurious senses while the average magnitude of the occurrence probability is unaffected.[6] Inclusion of the occurrence probability term will lead to the dominance of this term over the mutual information term, resulting in the system flavouring the sense with the more frequently occurring defining concepts most of the time.

2. The score of a sense with respect to the current context is normalised by subtracting the score of the sense calculated with respect to the GlobalCS (which contains all defining concepts) from it (see formula [1]). In effect, we are comparing the score between the sense with the current context and the score between the sense and an artificially constructed "average" context. This is needed to rectify the bias towards the sense(s) with defining concepts of higher average mutual information (over the set of all defining concepts), which is intensified by the ambiguity of the context words.

3. Negative mutual information score is taken to be 0 ([6]). Negative mutual information is unreliable due to the smaller number of data points.

4. The evidence (mutual information score) from multiple defining concepts/words is averaged rather than summed ([2], [4] & [5]). This is to compensate for the different lengths of definitions of different senses and different lengths of the context. The evidence from a polysemous context word is taken to be the evidence from its sense with the highest mutual information score ([3]). This is due to the fact that only one of the senses is used in the given sentence.

3 Evaluation

Our system is tested on the twelve words discussed in Yarowsky (1992) and previous publications on sense disambiguation. Results are shown in Table 1. Our system achieves an average accuracy of 77% on a mean 3-way sense distinction over the twelve words. Numerically, the result is not as good as the 92% as reported in Yarowsky (1992). However, direct comparison between the numerical results can be misleading since the experiments are carried out on two very different corpora both in size and genre. Firstly, Yarowsky's system is trained with the 10 million word Grolier's Encyclopedia, which is a magnitude larger than the Brown corpus used by our system. Secondly, and more importantly, the two corpora, which are also the test corpora, are very different in genre. Semantic coherence of text, on which both systems rely, is generally stronger in technical writing than in most other kinds of text. Statistical disambiguation systems which rely on semantic coherence will generally perform better on technical writing, which encyclopedia entry can be regarded as one kind of, than on most other kinds of text. On the other hand, the Brown corpus is a collection of text with all kinds of genre.

People make use of syntactic, semantic and pragmatic knowledge in sense disambiguation. It is not very realistic to expect any system which only possesses semantic coherence knowledge (including

[5] The occurrence probabilities of some defining concepts will not be independent in some contexts. However, modelling the dependency between different concepts in different contexts will lead to an explosion of the complexity of the model.

[6] The noise only leads to incorrect distribution of the occurrence probability.

ours as well as Yarowsky's) to achieve a very high level of accuracy for all words in general text. To provide a better evaluation of our approach, we have conducted an informal experiment aiming at establishing a more reasonable upper bound of the performance of such systems. In the experiment, a human subject is asked to perform the same disambiguation task as our system, given the same contextual information.[7] Since our system only uses semantic coherence information and has no deeper understanding of the meaning of the text, the human subject is asked to disambiguate the target word, given a list of all the content words in the context (sentence) of the target word in random order. The words are put in random order because the system does not make use of syntactic information of the sentence either. The human subject is also allowed access to a copy of LDOCE which the system also uses. The results are listed in Table 1. The actual upper bound of the performance of statistical methods using semantic coherence information only should be slightly better than the performance of human since the human is disadvantaged by a number of factors, including but not limited to: 1. it is unnatural for human to disambiguate in the described manner; 2. the semantic coherence knowledge used by the human is not complete or specific to the current corpus[8]; 3. human error. However, the results provide a rough approximation of the upper bound of performance of such systems.

The human subject achieves an average accuracy of 71% over the twelve words, which is 6% lower than our system. More interestingly, the results of the human subject are found to exhibit a similar pattern to the results of our system - the human subject performs better on words and senses for which our system achieve higher accuracy and less well on words and senses for which our system has a lower accuracy.

4 The Use of Sentence as Local Context

Another significant point our experiments have shown is that the sentence can also provide enough contextual information for semantic coherence based approaches in a large proportion of cases.[9] The average sentence length in the Brown corpus is 19.4[10] words which is 5 times smaller than the 100 word window used in Gale et al. (1992) and Yarowsky (1992). Our approach works well even with a small "window" because it is based on the identification of salient concepts rather than salient words. In salient word based approaches, due to the problem of data sparseness, many less frequently occurring words which are intuitively salient to a particular word sense will not be identified in practice unless an extremely large corpus is used. Therefore the sentence usually does not contain enough identified salient words to provide enough contextual information. Using conceptual co-occurrence data, contextual information from the salient but less frequently used words in the sentence will also be utilised through the salient concepts in the conceptual expansions of these words. Obviously, there are still cases where the sentence does not provide enough contextual information even using conceptual co-occurrence data, such as when the sentence is too short, and contextual information from a larger context has to be used. However, the ability to make use of information in a smaller context is very important because the smaller context always overrules the larger context if their sense preferences are different. For example, in a legal trial context, the correct sense of *sentence* in the clause *she was asked to repeat the last word of her previous sentence* will be its *word* sense rather than its *legal* sense which would have been selected if a larger context is used instead.

[7] The result is less than conclusive since only one human subject is tested. In order to acquire more reliable results, we are currently seeking a few more subjects to repeat the experiment.

[8] The subject has not read through the whole corpus.

[9] Analysis of the test samples which our system fails to correctly disambiguate also shows that increasing the window size will benefit the disambiguation process only in a very small proportion of these samples. The main cause of errors is the polysemous words in dictionary definitions which we will discuss in Section 6.

[10] Based on 1004998 words and 51763 sentences.

Table 1. Results of Experiments

Sense	N	DBCC	Human	Thes.
BASS				
Fish	1	100%	100%	100%
Musical senses	15	93%	100%	99%
	16	94%	100%	99%
BOW				
bending forward	1	0%	100%	--
weapon	0	--	--	92%
violin part	2	100%	100%	100%
knot	4	100%	100%	25%
front of ship	2	50%	100%	94%
bend in object *	--	--	--	50%
	9	78%	100%	91%
CONE				
shaped object	5	100%	100%	61%
fruit of a plant	0	--	--	99%
part of eye *	--	--	--	69%
	5	100%	100%	77%
DUTY				
obligation	54	57%	72%	96%
tax	2	100%	100%	96%
	56	59%	73%	96%
GALLEY				
ancient ship	0	--	--	97%
ship's kitchen	4	100%	50%	50%
printer's tray	0	--	--	100%
	4	100%	50%	95%
INTEREST				
curiosity	187	43%	41%	88%
advantage	59	42%	47%	34%
share	8	25%	38%	38%
money paid	48	88%	75%	90%
	302	49%	47%	72%
ISSUE				
bringing out	36	64%	75%	89%
important point	87	56%	40%	94%
stock *	--	--	--	100%
	123	59%	50%	94%
MOLE				
skin blemish	2	50%	50%	100%
animal	0	--	--	100%
stone wall **	1	100%	100%	--
quantity *	--	--	--	98%
machine *	--	--	--	100%
	3	67%	67%	99%
SENTENCE				
punishment	11	91%	100%	99%
group of words	20	80%	45%	98%
	31	84%	65%	98%

Sense	N	DBCC	Human	Thes.
SLUG				
animal	1	0%	0%	100%
fake coin	0	--	--	50%
type strip	0	--	--	100%
bullet	4	100%	50%	100%
mass unit *	--	--	--	100%
metallurgy *	--	--	--	100%
	5	80%	40%	97%
STAR				
space object	4	75%	75%	96%
shaped object	0	--	--	95%
celebrity	11	45%	64%	82%
	15	53%	67%	96%
TASTE				
flavour	21	100%	95%	93%
preference	26	96%	85%	93%
	47	98%	89%	93%

Notes:

1. *N* marks the column with the number of test samples for each sense. *DBCC* (Definition-Based Conceptual Co-occurrence) and *Human* mark the columns with the results of our system and the human subject in disambiguating the occurrences of the 12 words in the Brown corpus, respectively. *Thes.* (thesaurus) marks the column with the results of Yarowsky (1992) tested on the Grolier's Encyclopedia.

2. The "correct" sense of each test sample is chosen by hand disambiguation carried out by the author using the sentence as the context. A small proportion of test samples cannot be disambiguated within the given context and are excluded from the experiment.

3. The senses marked with * are used in Yarowsky (1992) but no corresponding sense is found in LDOCE.

4. The sense marked with ** is defined in LDOCE but not used in Yarowsky (1992).

6. In our experiment, the words are disambiguated between all the senses listed except the ones marked with *.

7. The rare senses listed in LDOCE are not listed here. For some of the words, more than one sense listed in LDOCE corresponds to a sense as used in Yarowsky (1992). In these cases, the senses used by Yarowsky are adopted for easier comparison.

8. All results are based on 100% recall.

5 Related Work

Previous attempts to tackle the data sparseness problem in general corpus-based work include the class-based approaches and similarity-based approaches. In these approaches, relationships between a given pair of words are modelled by analogy with other words that resemble the given pair in some way. The class-based approaches (Brown et al., 1992; Resnik, 1992; Pereira et al., 1993) calculate co-occurrence data of words belonging to different classes,[11] rather than individual words, to enhance the co-occurrence data collected and to cover words which have low occurrence frequencies. Dagan et al. (1993) argue that using a relatively small number of classes to model the similarity between words may lead to substantial loss of information. In the similarity-based approaches (Dagan et al., 1993 & 1994; Grishman et al., 1993), rather than a class, each word is modelled by its own set of *similar words* derived from statistical data collected from corpora. However, deriving these sets of similar words requires a substantial amount of statistical data and thus these approaches require relatively large corpora to start with.[12]

Our definition-based approach to statistical sense disambiguation is similar in spirit to the similarity-based approaches, with respect to the "specificity" of modelling individual words. However, using definitions from existing dictionaries rather than derived sets of similar words allows our method to work on corpora of much smaller sizes. In our approach, each word is modelled by its own set of defining concepts. Although only 1792 defining concepts are used, the set of all possible combinations (a power set of the defining concepts) is so huge that it is very unlikely two word senses will have the same combination of defining concepts unless they are almost identical in meaning. On the other hand, the thesaurus-based method of Yarowsky (1992) may suffer from loss of information (since it is semi-class-based) as well as data sparseness (since it is based on salient words) and may not perform as well on general text as our approach.

6 Limitation and Further work

Being a dictionary-based method, the natural limitation of our approach is the dictionary. The most serious problem is that many of the words in the controlled vocabulary of LDOCE are polysemous themselves. The result is that many of our list of 1792 defining concepts actually stand for a number of distinct concepts. For example, the defining concept *point* is used in its *place* sense, *idea* sense and *sharp end* sense in different definitions. This affects the accuracy of disambiguating senses which have definitions containing these polysemous words and is found to be the main cause of errors for most of the senses with below-average results.

We are currently working on ways to disambiguate the words in the dictionary definitions. One possible way is to apply the current method of disambiguation on the defining text of dictionary itself. The LDOCE defining text has roughly half a million words in its 41000 entries, which is half the size of the Brown corpus used in the current experiment. Although the result on the dictionary cannot be expected to be as good as the result on the Brown corpus due to the smaller size of the dictionary, the reliability of further co-occurrence data collected and, thus, the performance of the disambiguation system can be improved significantly as long as the disambiguation of the dictionary is considerably more accurate than by chance.

Our success in using definitions of word senses to overcome the data sparseness problem may also lead to further improvement of sense disambiguation technologies. In many cases, semantic coherence information is not adequate to select the correct sense, and knowledge about local constraints is needed.[13] For disambiguation of polysemous nouns, these constraints include the modifiers of these nouns and the verbs which take these nouns as objects, etc. This knowledge has been successfully acquired from corpora in manual or semi-automatic approaches such as that described in Hearst (1991). However, fully automatic lexically based approaches

---

[11] Classes used in Resnik (1992) are based on the WordNet taxonomy while classes of Brown et al. (1992) and Pereira et al. (1993) are derived from statistical data collected from corpora.

[12] The corpus used in Dagan et al. (1994) contains 40.5 million words.

[13] Hatzivassiloglou (1994) shows that the introduction of linguistic cues improves the performance of a statistical semantic knowledge acquisition system in the context of word grouping.

such as that described in Yarowsky (1992) are very unlikely to be capable of acquiring this finer knowledge because the problem of data sparseness becomes even more serious with the introduction of syntactic constraints. Our approach has overcome the data sparseness problem by using the defining concepts of words. It is found to be effective in acquiring semantic coherence knowledge from a relatively small corpus. It is possible that a similar approach based on dictionary definitions will be successful in acquiring knowledge of local constraints from a reasonably sized corpus.

7 Conclusion

We have shown that using definition-based conceptual co-occurrence data collected from a relatively small corpus, our sense disambiguation system has achieved accuracy comparable to human performance given the same amount of contextual information. By overcoming the data sparseness problem, contextual information from a smaller local context becomes sufficient for disambiguation in a large proportion of cases.

Acknowledgments

I would like to thank Robert Dale and Vance Gledhill for their helpful comments on earlier drafts of this paper, and Richard Buckland and Mark Dras for their help with the statistics.

References

Black, E., 1988. An Experiment In Computational Discrimination of English Word Senses. *IBM Journal of research and development*, vol. 32, pp.185-194.

Brown, P., et al., 1991. Word-sense Disambiguation using Statistical Methods. In *Proceedings of 29th annual meeting of ACL*, pp.264-270.

Brown, P. et al., 1992. Class-based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467-479.

Church, K. and P. Hanks, 1989. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pp.76-83.

Dagan, I. et al., 1991. Two Languages Are More Informative Than One. In *Proceedings of the 29th Annual Meeting of the ACL*, pp130-137.

Dagan, I. et al., 1993. Contextual Word Similarity and Estimation From Sparse Data. In *Proceedings of the 31st Annual Meeting of the ACL.*

Dagan, I. et al., 1994. Similarity-Based Estimation of Word Cooccurrence Probabilities. In *Proceedings of the 32nd Annual Meeting of the ACL*, Las Cruces, pp272-278.

Fano, R., 1961. Transmission of Information. MIT Press, Cambridge, Mass.

Gale, W., et al., 1992. A Method for Disambiguating Word Senses in a Large Corpus. *Computer and Humanities*, vol. 26 pp.415-439.

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

Hatzivassiloglou, V., 1994. Do We Need Linguistics When We Have Statistics? A Comparative Analysis of the Contributions of Linguistic Cues to a Statistical Word Grouping System. In *Proceedings of Workshop The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, Las Cruces, New Mexico. Association of Computational Linguistics.

Hearst, M., 1991. Noun Homograph Disambiguation Using Local Context in Large Text Corpora, *Using Corpora*, University of Waterloo, Waterloo, Ontario.

Kelly, E. and P. Stone, 1975. *Computer Recognition of English Word Senses*, North-Holland, Amsterdam.

Pereira F., et al., 1993. Distributional Clustering of English words. In Proceedings of the 31st Annual Meeting of the ACL. pp183-190.

Procter, P., et al. (eds.), 1978. *Longman Dictionary of Contemporary English*, Longman Group.

Resnik, P., 1992. WordNet and distributional analysis: A class-based approach to lexical discovery. In Proceedings of AAAI Workshop on Statistically-based NLP Techniques, San Jose, California.

Yarowsky, D., 1992. Word-sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings of COLING92*, pp.454-460.