Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora

David Yarowsky

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ, 07974
yarowsky@research.att.com

Abstract

This paper describes a program that disambiguates English word senses in unrestricted text using statistical models of the major Roget's Thesaurus categories. Roget's categories serve as approximations of conceptual classes. The categories listed for a word in Roget's index tend to correspond to sense distinctions; thus selecting the most likely category provides a useful level of sense disambiguation. The selection of categories is accomplished by identifying and weighting words that are indicative of each category when seen in context, using a Bayesian theoretical framework.

Other statistical approaches have required special corpora or hand-labeled training examples for much of the lexicon. Our use of class models overcomes this knowledge acquisition bottleneck, enabling training on unrestricted monolingual text without human intervention. Applied to the 10 million word Grolier's Encyclopedia, the system correctly disambiguated 92% of the instances of 12 polysemous words that have been previously studied in the literature.

1 Problem Formulation

This paper presents an approach to word sense disambiguation that uses classes of words to derive models useful for disambiguating individual words in context. "Sense" is not a well defined concept; it has been based on subjective and often subtle distinctions in topic, register, dialect, collocation, part of speech and valency. For the purposes of this study, we will define the senses of a word as the categories listed for that word in Roget's International Thesaurus (Fourth Edition - Chapman, 1977).[1] *Sense disambiguation* will constitute selecting the listed category which is most probable given the surrounding context. This may appear to be a particularly crude approximation, but as shown in the example below and in the table of results, it is surprisingly successful.

Label	**Training Contexts**
SENSE-1	Treadmills attached to **cranes** were used to lift heavy
SENSE-1	for supplying power for **cranes** , hoists , and lifts .
SENSE-1	bove this height , a tower **crane** is often used .SB This
SENSE-2	elaborate courtship rituals **cranes** build a nest of vegetati
SENSE-2	are more closely related to **cranes** and rails .SB They ran
SENSE-2	low trees .PP At least five **crane** species are in danger of

[1] Note that this edition of Roget's Thesaurus is much more extensive than the 1911 version, though somewhat more difficult to obtain in electronic form. One could use other other concept hierarchies, such as *WordNet* (Miller, 1990) or the LDOCE subject codes (Slator, 1991). All that is necessary is a set of semantic categories and a list of the words in each category.

Not only do the Roget categories succeed in partitioning the major senses, but the sense tags they provide as output are far more mnemonic than a dictionary numbering such as "crane 1.2". Should such a dictionary sense number be desired as output, section 5 will outline how a linkage between Roget categories and dictionary definitions can be made.

We will also focus on sense distinctions *within* a given part of speech. Distinctions *between* parts of speech, should be based on local syntactic evidence. We use a stochastic part-of-speech tagger (Church, 1989) for this purpose, run as a preprocessor.

2 Proposed Method

The strategy proposed here is based on the following three observations: 1) Different conceptual classes of words, such as ANIMALS or MACHINES tend to appear in recognizably different contexts. 2) Different word senses tend to belong to different conceptual classes (*crane*) can be an ANIMAL or a MACHINE). 3) If one can build a context discriminator for the conceptual classes, one has effectively built a context discriminator for the word senses that are members of those classes. Furthermore, the context indicators for a Roget category (e.g. *gear*, *piston* and *engine* for the category TOOLS/MACHINERY) will also tend to be context indicators for the members of that category (such as the machinery sense of *crane*).

We attempt to identify, weight and utilize these indicative words as follows. For each of the 1042 Roget Categories:

1. Collect contexts which are representative of the Roget category
2. Identify salient words in the collective context and determine weights for each word, and
3. Use the resulting weights to predict the appropriate category for a polysemous word occurring in novel text.

Step 1: Collect Contexts which are Representative of the Roget category

The goal of this step is to collect a set of words that are typically found in the context of a Roget category. To do this, we extract *concordances* of 100 surrounding words for each occurrence of each member of the category in the corpus. Below is a sample set of partial concordances for words in the category TOOLS/MACHINERY (348). The complete set contains 30,924 lines, selected from the particular training corpus used in this study, the 10 million word, June 1991 electronic version of Grolier's Encyclopedia.

Training Data (Words in Context)
... CARVING .SB The gutter **adz** has a concave blade for form ...
... uipment such as a hydraulic **shovel** capable of lifting 26 cubic ...
... on .SB Resembling a power **shovel** mounted on a floating hul ...
... uipment , valves for nuclear **generators** , oil-refinery turbines ...
... 00 BC , flint-edged wooden **sickles** were used to gather wild ...
... l-penetrating carbide-tipped **drills** forced manufacturers to fi ...
... ent heightens the colors .SB **Drills** live in the forests of equa ...
... traditional ABC method and **drill** were unchanged , and dissa ...
... nter of rotation .PP A tower **crane** is an assembly of fabricat ...
... rshy areas .SB The crowned **crane** , however , occasionally ...

For optimal training, the concordance set should only include references to the given category. But in practice it will unavoidably include spurious examples since many of the words are polysemous (such as *drill* and *crane* in lines 7, 8, and 10 above).

While the level of noise introduced through polysemy is substantial, it can usually be tolerated because the spurious senses are distributed through the 1041 other categories, whereas the signal is concentrated in just one. Only if several words had secondary senses in the same category would context typical for the other category appear significant in this context.

However, if one of these spurious senses was frequent and dominated the set of examples, the situation could be disastrous. An attempt is made to weight the concordance data to minimize this effect and to make the sample representative of all tools and machinery, not just the more common ones. If a word such as *drill* occurs k times in the corpus, all words in the context of *drill* contribute weight $1/k$ to frequency sums.

Despite its flaws, this weighted matrix will serve as a representative, albeit noisy, sample of the typical context of TOOLS/MACHINERY in Grolier's encyclopedia.

Step 2: Identify salient words in the collective context, and weight appropriately

Intuitively, a salient word[2] is one which appears significantly more often in the context of a category than at other points in the corpus, and hence is a better than average indicator for the category. We formalize this with a mutual-information-like estimate: $\frac{Pr(w|RCat)}{Pr(w)}$, the probability of a word (w) appearing in the context of a Roget category divided by its overall probability in the corpus.

It is important to exercise some care in estimating $Pr(w|RCat)$. In principle, one could simply count the number of times that w appears in the collective context. However, this estimate, which is known as the maximum likelihood estimate (MLE), can be unreliable, especially when w does not appear very often in the collective context. We have smoothed the local estimates of $Pr(w|RCat)$ with global estimates of $Pr(w)$ to obtain a more reliable estimate. Estimates obtained from the local context are subject to measurement errors whereas estimates obtained from the global context are subject to being irrelevant. By interpolating between the two, we attempt to find a compromise between the two sources of error. This procedure is based on recent work pioneered by William Gale, and is explained in detail in another paper (Gale, Church and Yarowsky, 1992). Space does not permit a complete description here.

Below are salient words for Roget categories 348 and 414. Those selected are the most *important* to the models, where *importance* is defined as the product of salience and local frequency. That is to say important words are distinctive *and* frequent.

The numbers in parentheses are the log of the salience ($log\frac{Pr(w|RCat)}{Pr(w)}$), which we will henceforth refer to as the word's *weight* in the statistical model of the category.

ANIMAL/INSECT (Category 414):
species (2.3), family (1.7), bird (2.6), fish (2.4), breed (2.2), cm (2.2), animal (1.7), tail (2.7), egg (2.2), wild (2.6), common (1.3), coat (2.5), female (2.0), inhabit (2.2), eat (2.2), nest (2.5),...

TOOLS/MACHINERY (Category 348):
tool (3.1), machine (2.7), engine (2.6), blade (3.8), cut (2.6), saw (5.1), lever (4.1), pump (3.5), device (2.2), gear (3.5), nife(3.8), wheel (2.8), shaft(3.3), wood(2.0), tooth(2.5), piston(3.6),...

Notice that these are not a list of members of the category; they are the words which are likely to co-occur with the members of the category. The complete list for TOOLS/MACHINERY includes

[2] For illustrative simplicity, we will refer to *words* in context. In practice, all operations are actually performed on the *lemmas* of the words (eat/V = eat,eats,eating,ate,eaten), and inflectional distinctions are ignored. While this achieves more concentrated and better estimated statistics, it throws away useful information which may be exploited in future work.

a broad set of relations, such as meronomy (*blade, engine, gear, wheel, shaft, tooth, piston* and *cylinder*), typical functions of machines (*cut, rotate, move, turn, pull*), typical objects of those actions (*wood, metal*), as well as typical modifiers for machines (*electric, mechanical, pneumatic*). The list for a category typically contains over 3000 words, and is far richer than can be derived from a dictionary definition.

Step 3: Use the resulting weights to predict the appropriate category for a word in novel text

When any of the salient words derived in step 2 appear in the context of an ambiguous word, there is evidence that the word belongs to the indicated category. If several such words appear, the evidence is compounded. Using Bayes' rule, we sum their weights, over all words in context, and determine the category for which the sum is greatest.[3]

$$\underset{RCat}{ARGMAX} \quad \sum_{w\ in\ context} \log(\frac{Pr(w|RCat) \times Pr(RCat)}{Pr(w)})$$

The context is defined to extend 50 words to the left and 50 words to the right of the polysemous word. This range was shown by Gale, Church and Yarowsky (1992) to be useful for this type of broad topic classification, in contrast to the relatively narrow (±3-6 word) window used in previous studies (e.g. Black, 1988). The maximization over RCats is constrained to consider only those categories under which the polysemous word is listed, generally on the order of a half dozen or so.[4]

For example the word *crane* appears 74 times in Groliers; 36 occurrences refer to the animal sense and 38 refer to the heavy machinery sense. The system correctly classified all but one of the machinery senses, yielding 99% overall accuracy . The one misclassified case had a low score for all models, indicating a lack of confidence in any classification.

It is useful to look at one example in some more detail. Consider the following instance of *crane* and its context of ±10 words:[5]

> ... lift water and to grind grain .PP Treadmills
> attached to **cranes** were used to lift heavy
> objects from Roman times , ...

The table below shows the strongest indicators identified for the two categories in the sentence above. The model weights, as noted above, are equivalent to $log\frac{Pr(w|RCat)}{Pr(w)}$. Several indicators were found for the TOOLS/MACHINE class. There is very little evidence for the ANIMAL sense of *crane*, with the possible exception of *water*. The preponderance of evidence favors the former classification, which happens to be correct. The difference between the two total scores indicate strong confidence in the answer.

[3] The reader may have noticed that the $Pr(w)$ factor can be omitted since it will not change the results of the maximization. It is included here for expository convenience so that it is possible to compare results across words with very different probabilities. The factor also becomes important when an incomplete set of indicators is stored because of computational space constraints. Currently we assume a uniform prior probability for each Roget category ($Pr(Rcat)$), i.e. sense classification is based exclusively on contextual information, independent of the underlying probability of a given Roget category appearing at any point in the corpus.

[4] Although it is often useful to restrict the search in this way, the restriction does sometimes lead to trouble, especially when there are gaps in the thesaurus. For example, the category AMUSEMENT (# 876) lists a number of card playing terms, but for some reason, the word *suit* is not included in this list. As it happens, the Grolier's Encyclopedia contains 54 instances of the card-playing sense of *suit*, all of which are mislabeled if the search is limited to just those categories of *suit* that are listed in Roget's. However, if we open up the search to consider all 1042 categories, then we find that all 54 instances of *suit* are correctly labeled as AMUSEMENT, and moreover, the score is large in all 54 instances, indicating great confidence in the assignment. It is possible that the unrestricted search mode might be a good way to attempt to fill in omissions in the thesaurus. In any case, when *suit* is added to the AMUSEMENT category, overall accuracy improves from 68% to 92%.

[5] This narrower window is used for illustrative simplicity.

TOOLS/MACHINE	Weight	ANIMAL/INSECT	Weight
lift	2.44	water	0.76
lift	2.44		
grain	1.68		
used	1.32		
heavy	1.28		
Treadmills	1.16		
attached	0.58		
grind	0.29		
water	0.11		
TOTAL	11.30	TOTAL	0.76

3 Evaluation

The algorithm described above was applied to 12 polysemous words previously discussed in the sense disambiguation literature. Table 1 shows the system's performance.[6]

Authors who have discussed these words are listed in the table in parentheses, along with the reported accuracy of their systems. Direct comparisons of performance between researchers is difficult, compounded by variances in corpora and grading criteria; using the same words is an attempt to minimize these differences.

Regrettably, most authors have reported their results in qualitative terms. The exceptions include Zernik (1990) who cited "recall and precision of over 70%" for one word *interest*) and observed that results for other words, including *issue*, were "less positive." Clear (1989) reported results for two words (65% and 67%), apparently at 85% recall. Lesk (1986) claimed overall "50-70%" accuracies, although it is unclear under which parameters and constraints. In a 5 word test set, Black (1988) observed 75% mean accuracy using his optimal method on high entropy, 4-way sense distinctions. Hearst (1991) achieved 84% on simpler 2-way distinctions, editing out additional senses from the test set. Gale, Church and Yarowsky (1992) reported 92% accuracy, also on 2-way distinctions.

Our current work compares favorably with these results, with 92% accuracy on a mean 3-way sense distinction.[7] The performance is especially promising given that no hand tagging or special corpora were required in training, unlike all other systems considered.

[6] 1) *Freq* refers to the total number of each sense observed in the test corpus. *Accuracy* indicates the percentage of those tagged correctly.

2) Because there is no independent ground truth to indicate which is the "correct" Roget category for a given word, the decision is a subjective judgement made by a single human judge, in this case the author.

3) As previously noted, the Roget index is incomplete. In four cases, identified by a $*$, one missing category has been added to the list of possibilities for a word. These omissions in the lexicon have been identified as outlined in Footnote 4. Without these additions, overall system performance would decrease by 5

4) Uses which an English speaker may consider a single sense are often realized by several Roget categories. For the purposes of succinct representation, such categories have been merged, and the name of the dominant category used in the table. As of this writing, the process has not been fully automated.

For many applications such as speech synthesis and assignment to an established dictionary sense number or possible French translations, this merging of Roget classes is not necessary.

The primary criterion for success is that words are partitioned into pure sense clusters. Words having a different sense from the majority sense of a partition are graded as errors.

[7] This result is a fair measure of performance on words used in previous studies, and may be useful for comparison across systems. However, as words previously discussed in the literature may not be representative of typical English polysemy, mean performance on a completely random set of words should differ.

				Accuracy	Total
Word	Sense	Roget Category	Freq	Per Sense	Accuracy
star	space object	UNIVERSE	1422	96%	96%
	celebrity	ENTERTAINER	222	95%	
	star-shaped object	INSIGNIA	56	82%	
mole	quantity	CHEMICALS	95	98%	99%
	mammal	ANIMAL	46	100%	
	skin blemish	DISEASE	13	100%	
	digging machine	SUPPORT	4	100%	
galley	ancient ship	SHIP,BOAT	35	97%	95%
	printer's tray	PRINTING	5	100%	
	ship's kitchen	COOKING	2	50%	
cone	part of tree	PLANT	71	99%	77%
	shape of object	ANGULARITY	89	61%	
	part of eye	VISION	13	69%	
bass	musical senses	MUSIC	158	99%	99%
	fish	ANIMAL	69	100%	
bow	weapon	ARMS	59	92%	91%
	front of ship	SHIP,BOAT	34	94%	
	violin part	MUSICAL_INSTR	30	100%	
	ribbon	ORNAMENTATION	4	25%	
	bend in object	CONVEXITY	2	50%	
	lowering head	RESPECT	0	–	
taste	preference	PARTICULARITY	228	93%	93%
	flavor	SENSATION	80	93%	
interest	curiosity	REASONING	359	88%	72%
	advantage	INJUSTICE	163	34%	
	financial	DEBT	59	90%	
	share	PROPERTY	21	38%	
issue	topic	POLITICS	831	94%	94%
	periodical	BOOKS,PERIOD	28	89%	
	stock	SECURITIES	9	100%	
duty	obligation	DUTY	347	96%	96%
	tax	PRICE,FEE	52	96%	
sentence	punishment	LEGAL_ACTION	128	99%	98%
	set of words	GRAMMAR	213	98%	
slug	animal	ANIMAL	24	100%	97%
	type strip	PRINTING	8	100%	
	mass unit	WEIGHT	3	100%	
	fake coin	MONEY	2	50%	
	metallurgy	IMPULSE,IMPACT	1	100%	
	bullet	ARMS	1	100%	

Table 1: Summary of performance

4 Limitations of the Method

The procedure described here is based on broad context models. It performs best on words with senses which can be distinguished by their broad context. These are most typically concrete nouns. Performance is weaker on the following:

Topic Independent Distinctions: One of the reasons that *interest* is disambiguated poorly is that it can appear in almost any context. While its "curiosity" sense is often indicated by the presence of an academic subject or hobbie, the "advantage" sense (to be in one's interests) has few topic constraints. Distinguishing between two such abstractions is difficult.[8] However, the financial sense of *interest* is readily identifiable, and can be distinguished from the non-financial uses with 92% accuracy. Other distinctions between topic independent and topic constrained senses appear successful as well (e.g. *taste, issue, duty* and *sentence*).

Minor Sense Distinctions within a Category: Distinctions between the medicinal and narcotic senses of *drug* are not captured by the system because they both belong to the same Roget category (REMEDY). Similar problems occur with the musical senses of *bass*. Roget's Thesaurus offers a rich sub-hierarchy within each category, however. Future implementations will likely use this information, which is currently ignored.

Verbs: Verbs have not been considered in this particular study, and it appears that they may benefit from more local models of their typical arguments. The unmodified system does seem to perform well on verbs which show clear topic distinctions such as *fire*. It's weapon, engine, furnace, employee, imagination and pottery senses have been disambiguated with 85% accuracy.

Pre-Nominal Modifiers: The disambiguation of pre-nominal modifiers (adjectives and compound nominals) is heavily dependent on the noun modified, and much less so on distant context. While class-based Bayesian discrimination may be useful here as well, the optimal window size is much narrower.

Idioms: These broad context, topic-based discriminators are also less successful in dealing with a word like *hand*, which is usually found in fixed expressions such as *on the other hand* and *close at hand*. These fixed expressions have more function than content, and therefore, they do not lend themselves to a method that depends on differences in content. The situation is far from hopeless, as many idioms are listed directly in Roget's Thesaurus and can be associated with a category through simple table lookup. Other research, such as Smadja and McKeown (1990), have shown more general ways of identifying and handling these fixed expressions and collocations.

Given the broad set of issues involved in sense disambiguation, it is reasonable to use several specialized tools in cooperation. We already handle part of speech distinctions through other methods; an efficient idiom recognizer would be an appropriate addition as well.

5 Linking Roget Categories with other Sense Representations

The Roget category names tend to be highly mnemonic and may well suffice as sense tags. However, one may want to link the Roget tags with an established reference such as the sense numbers one finds in a dictionary. We accomplish this by applying the models described above to the text of the definitions in a dictionary, creating a table of correspondences between Roget categories and sense numbers. Results for the word *crane* are illustrated below for two dictionaries: (1) COBUILD (Sinclair, 1987), and (2) Collins English Dictionary, First Edition (CED1) (Hanks, 1979).

[8] Black (1988) has noted that this distinction for *interest* is strongly correlated with the plurality of the word, a feature we currently don't utilize.

Dictionary	RCAT	Sense #	Definition
COBUILD	MACHINE	crane 1.1	a machine with a long movable
	ANIMAL	crane 1.2	large bird with a long neck and
Collins English	ANIMAL	crane 1	a large long-necked long-leg
	ANIMAL	crane 2	any similar bird , such as a her
	MACHINE	crane 3	a device for lifting and moving
	MACHINE	crane 4	a large trolley carrying a boom

It may also be possible to link Roget category tags with "natural" sense tags, such as translations in a foreign language. We use a word-aligned parallel bilingual corpus such as the French-English Canadian Hansards for this purpose. For example, consider the polysemous word *duty* which can be translated into French as *devoir* or *droit*, depending on the sense (obligation or tax, respectively). When the Grolier-trained models are applied to the English side of the Hansards, the words tagged PRICE,FEE most commonly aligned with the French words *droits* (256), *droit* (96) and *douane* (67). Words labeled DUTY (the Roget category for Obligation) most frequently aligned with *devoir* (205). These correlations may have useful implications for machine translation and bilingual lexicography.

6 Other Sense Disambiguation Methods: The Knowledge Acquisition Bottleneck

Word sense disambiguation is a long-standing problem in computational linguistics (Kaplan, 1950; Yngve, 1955; Bar-Hillel, 1960), with important implications for a variety of practical applications including speech synthesis, information retrieval, and machine translation. Most approaches may be characterized by the following generalizations: 1) They tend to focus on the search for sets of word-specific features or indicators (typically words in context) which can disambiguate the senses of a word. 2) Efforts to acquire these indicators have faced a knowledge acquisition bottleneck, characterized by either substantial human involvement for each word, and/or incomplete vocabulary coverage.

The AI community has enjoyed some success hand-coding detailed "word experts" (Small and Rieger, 1982; Hirst, 1987), but this labor intensive process has severely limited coverage beyond small vocabularies.

Others such as Lesk (1986), Walker (1987), Veronis and Ide (1990), and Guthrie et al. (1991) have turned to machine readable dictionaries (MRD's) in an effort to achieve broad vocabulary coverage. MRD's have the useful property that some indicative words for each sense are directly available in numbered definitions and examples. However, definitions are often too short to provide an adequate set of indicators, and those words which are found lack significance weights to identify which are crucial and which are merely chaff. Dictionaries provide well structured but incomplete information.

Recently, many have turned to text corpora to broaden the range and volume of available examples. Unlike dictionaries, however, raw corpora do not indicate which sense of a word occurs at a given instance. Several researchers (Kelly and Stone, 1975; Black, 1988) have overcome this through hand tagging of training examples, and were able to discover useful discriminatory patterns from the partitioned contexts. This also has proved labor intensive. Others (Weiss, 1973; Zernik, 1990; Hearst, 1991) have attempted to partially automate the hand-tagging process through bootstrapping. Yet this has still required significant human intervention for each word in the vocabulary.

Brown et al. (1991), Dagan (1991), and Gale et al. (1992) have looked to parallel bilingual corpora to further automate training set acquisition. By identifying word correspondences in a bilingual text such as the Canadian Parliamentary Proceedings (Hansards), the translations found for each English word may serve as sense tags. For example, the senses of *sentence* may be identified through their correspondence in the French to *phrase* (grammatical sentence) or *peine* (legal sentence). While this method has been used successfully on a portion of the vocabulary, its coverage is also limited. Currently available bilingual corpora lack size or diversity: over half of the words considered in this study either never appear in the Hansards or lack examples of secondary senses. More fundamentally, many words are mutually ambiguous across languages. French would be of little use in disambiguating the word *interest*, as all major senses translate as *intérêt*. More promising is a non-Indo European language such as Japanese, which should avoid such mutual ambiguity for etymological reasons. Until more diverse, large bilingual corpora become available, the coverage of these methods will remain limited.

Each of these approaches have faced a fundamental obstacle: *word sense* is an abstract concept that is not identified in natural text. Hence any system which hopes to acquire discriminators for specific senses of a word will need to isolate samples of those senses. While this process has been partially automated, it appears to require substantial human intervention to handle an unrestricted vocabulary.

7 Conclusion

This paper has described an approach to word sense disambiguation using statistical models of word classes. This method overcomes the knowledge acquisition bottleneck faced by word-specific sense discriminators. By entirely circumventing the issue of polysemy resolution in training material acquisition, the system has acquired an extensive set of sense discriminators from unrestricted monolingual texts without human intervention. Class models also offer the additional advantages of smaller model storage requirements and increased implementation efficiency due to reduced dimensionality. Also, they can correctly identify a word sense which occurs rarely or only once in the corpus – performance unattainable by statistically trained word-specific models. These advances are not without cost, as class-based models have diluted discriminating power and may not capture highly indicative collocations specific to only one word. Despite the inherent handicaps, the system performs better than several previous approaches, based on a direct comparison of results for the same words.

Acknowledgements

Special thanks are due to Ken Church and Barbara Grosz for their invaluable help in restructuring this paper, and to Bill Gale for the theoretical foundations on which this work rests. The author is also grateful to Marti Hearst, Fernando Pereira, Don Hindle, Richard Sproat, and Michael Riley for their comments and suggestions.

References

[1] Y. Bar-Hillel. Automatic translation of languages. In *Advances in Computers*, Donald Booth and R. E. Meagher, eds., Academic Press, New York, 1960.

[2] E. Black. An experiment in computational discrimination of English word senses. In *IBM Journal of Research and Development*, 232:185–194, 1988.

[3] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. Word sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 264–270, Berkeley, 1991.

[4] P. Brown, V. Della Pietra, S. Della Pietra, and R. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 1992.

[5] R. Chapman. *Roget's International Thesaurus (Fourth Edition).* Harper and Row, New York, 1977.

[6] Y. Choueka and S. Lusignam. Disambiguation by short contexts. *Computers and the Humanities*, 19:147-158, 1985.

[7] K. W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 136–143, 1988.

[8] J. Clear. An experiment in automatic word sense identification. Internal Document, Oxford University Press, Oxford, 1989.

[9] G. Cottrell. *A Connectionist Approach to Word Sense Disambiguation.* Pitman, London, 1989.

[10] I. Dagan, A. Itai, and U. Schwall. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 130-137, 1991.

[11] W. Gale, K. Church, and D. Yarowsky. Discrimination decisions for 100,000-dimensional spaces. *AT&T Statistical Research Report No. 103.*

[12] W. Gale, K. Church, and D. Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439, 1992.

[13] R. Granger. *FOUL-UP*: A program that figures out meanings of words from context. In *Proceedings, IJCAII-77*, pp. 172-178, 1977.

[14] J. Guthrie, L. Guthrie, Y. Wilks and H. Aidinejad. Subject dependent co-occurrence and word sense disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 146–152, 1991.

[15] P. Hanks (ed.). *Collins English Dictionary.* Collins, London and Glasgow, 1979.

[16] M. Hearst. Noun homograph disambiguation using local context in large text corpora. In *Using Corpora*, University of Waterloo, Waterloo, Ontario, 1991.

[17] G. Hirst. *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press, Cambridge, 1987.

[18] A. Kaplan. An experimental study of ambiguity in context. Cited in *Mechanical Translation*, 1, 1-3, 1950.

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

[20] M. Lesk. Automatic sense disambiguation: how to tell a pine cone from an ice cream cone. In *Proceeding of the 1986 SIGDOC Conference*, Association for Computing Machinery, New York, 1986.

[21] G. Miller. WordNet: an on-line lexical database. *International Journal of Lexicography*, 3, 4, 1990.

[22] F. Mosteller and D. Wallace. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, Massachusetts, 1964.

[23] G. Salton. *Automatic Text Processing.* Addison-Wesley Publishing Co, 1989.

[24] B. Slator. Using context for sense preference. In *Text-Based Intelligent Systems: Current Research in Text Analysis, Information Extraction and Retrieval*, P.S. Jacobs, ed., GE Research and Development Center, Schenectady, New York, 1990.

[25] F. Smadja, and K. McKeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 1990.

[26] S. Small, and C. Rieger. Parsing and comprehending with word experts (a theory and its realization). In *Strategies for Natural Language Processing*, W. Lehnert and M. Ringle. eds., Lawrence Erlbaum Associates, Hillsdale, NJ, 1982.

[27] J. Veronis and N. Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings, COLING-90*, pp. 389–394, 1990.

[28] D. Walker. Knowledge resource tools for accessing large text files. In *Machine Translation: Theoretical and Methodological Issues.* Sergei Nirenberg, ed., Cambridge University Press, Cambridge, England, 1987.

[29] S. Weiss. Learning to disambiguate. *Information Storage and Retrieval*, 9:33-41, 1973.

[30] V. Yngve. Syntax and the problem of multiple meaning. In *Machine Translation of Languages*, W. Locke and D. Booth, eds. Wiley, New York, 1955.

[31] U. Zernik. Tagging word senses in a corpus: the needle in the haystack revisited. In *Text-Based Intelligent Systems: Current Research in Text Analysis, Information Extraction and Retrieval*, P.S. Jacobs, ed., GE Research and Development Center, Schenectady, New York, 1990.