EXPERIMENTS IN SYNTACTIC AND SEMANTIC CLASSIFICATION AND DISAMBIGUATION USING BOOTSTRAPPING*

(Assoc. for Comp'l Linguistics Workshop on Acquisition of Lexical Knowledge from Text)
(June 1993, 117-127)

Robert P. Futrelle and Susan Gauch

Biological Knowledge Laboratory
College of Computer Science
Northeastern University
Boston, MA 02115

Internet: {futrelle, sgauch}@ccs.neu.edu

ABSTRACT

Methods that generate word classes without requiring pretagging have had notable success in the last few years (bootstrap methods or unsupervised classification). The methods described here strengthen these approaches and produce excellent word classes from a 200,000 word corpus. The method uses mutual information measures plus positional information from the words in the immediate context of a target word to compute similarities. Using the similarities, classes are built using hierarchical agglomerative clustering. At the leaves of the classification tree, words are grouped by syntactic and semantic similarity. Further up the tree, the classes are primarily syntactic. Once the initial classes are found they can be used to improve the classification of single word instances — to do classic word tagging. This is done by expanding each context word of a target instance into a tightly defined class of similar words, a *simset*. The use of simsets is shown to increase the tagging accuracy from 83% to 92% for the forms "cloned" and "deduced".

INTRODUCTION

The identification of the syntactic class and semantic information for words not contained in any on-line dictionary or thesaurus is an important and challenging problem. Excellent methods have been developed for part-of-speech (POS) tagging using stochastic models trained on partially tagged corpora (Church, 1988; Cutting, Kupiec, Pedersen, & Sibun, 1992). Semantic issues have been addressed, particularly for sense disambiguation, by using large contexts, e.g., 50 nearby words (Gale, Church, & Yarowsky, 1992) or by reference to on-line dictionaries (Krovetz, 1991; Lesk, 1986; Liddy & Paik, 1992; Zernik, 1991). More recently, methods to work with entirely untagged corpora have been developed which show great promise (Brill & Marcus, 1992; Finch & Chater, 1992; Myaeng & Li, 1992; Schutze, 1992). They are particularly useful for text with specialized vocabularies and word-use. These methods of unsupervised classification typically have clustering algorithms at their heart (Jain & Dubes, 1988). They use similarity of contexts (the distribution principle) as a measure of distance in the space of words and then cluster similar words into classes. This paper demonstrates a particular approach to these classification techniques.

---

* This material is based upon work supported by the National Science Foundation under Grant No. DIR-8814522.

In our approach, we take into account both the relative positions of the nearby context words as well as the mutual information (Church & Hanks, 1990) associated with the occurrence of a particular context word. The similarities computed from these measures of the context contain information about both syntactic and semantic relations. For example, high similarity values are obtained for the the two semantically similar nouns, "diameter" and "length" , as well as for the two adjectives "nonmotile" and "nonchemotactic".

We demonstrate the technique on three problems, all using a 200,000 word corpus composed of 1700 abstracts from a specialized field of biology: #1: Generating the full classification tree for the 1,000 most frequent words (covering 80% of all word occurrences). #2: The classification of 138 occurrences of the -*ed* forms, "cloned" and "deduced" into four syntactic categories, including improvements by using expanded context information derived in #1. #3: The classification of 100 words that only occur once in the entire corpus (*hapax legomena*), again using expanded contexts.

The results described below were obtained using no pretagging or on-line dictionary, but the results compare favorably with methods that do. The results are discussed in terms of the semantic fields they delineate, the accuracy of the classifications and the nature of the errors that occur. The results make it clear that this new technology is very promising and should be pursued vigorously. The power of the approach appears to result from using a focused corpus, using detailed positional information, using mutual information vectors and using a clustering method that updates the detailed context information when each new cluster is formed. Our approach was inspired by the fascinating results achieved by Finch and Chater at Edinburgh and the methods they used (Finch & Chater, 1992).

THE CORPUS — TECHNICAL, FOCUSED AND SMALL

In the Biological Knowledge Laboratory we are pursuing a number of projects to analyze, store and retrieve biological research papers, including working with full text and graphics (Futrelle, Kakadiaris, Alexander, Carriero, Nikolakis, & Futrelle, 1992; Gauch & Futrelle, 1993). The work is focused on the biological field of bacterial chemotaxis. A biologist has selected approximately 1,700 documents representing all the work done in this field since its inception in 1965. Our study uses the titles for all these documents plus all the abstracts available for them. The resulting corpus contains 227,408 words with 13,309 distinct word forms, including 5,833 words of frequency 1. There are 1,686 titles plus 8,530 sentences in the corpus. The sentence identification algorithm requires two factors — contiguous punctuation (".", "!", or "?") and capitalization of the following token. To eliminate abbreviations, the token prior to the punctuation must not be a single capital letter and the capitalized token after the punctuation may not itself be followed by a contiguous ".".

An example of a sentence from the corpus is,

> "$pre2$ $pre1$ one of the open reading frames was translated into a protein with pct amino acid identity to S. typhimurium FliI and pct identity to the beta subunit of E. coli ATP synthase $pos1$ $pos2$"

The positional items $pre... and $pos... have been added to furnish explicit context for sentence initial and sentence final constituents. Numbers have been converted to three forms corresponding to integers, reals and percentages ("pct in the example above). The machine-readable version of the corpus uses double quoted items to ease processing by Lisp, our language of choice.

The terminology we will use for describing words is as follows:

- **Target word:** A word to be classified.

- **Context words:** Appearing within some distance of a target word, "The big brown cat on the mat...".

- **Word class:** Any defined set of word forms or labeled instances.

- **Simset:** A word class in which each item, an *expansion word*, has a similarity greater than some chosen cutoff to a single *base word*.

- **Labeled instances:** Forms such as "cloned48" or "cloned73VBN", that would replace an occurrence of "cloned".

DESCRIBING AND QUANTIFYING WORD CONTEXTS

In these experiments, the context of a target word is described by the preceding two context words and the following two context words, Figure 1. Each position is represented by a vector corresponding to the occurrence of the 150 highest frequency words in the corpus, giving a 600-dimensional vector describing the context. Initially, the counts from all instances of a target word form w_i are summed so that the entry in the corresponding context word position in the vector is the sum of the occurrences of that context word in that position for the corresponding target word form; it is the joint frequency of the context word.

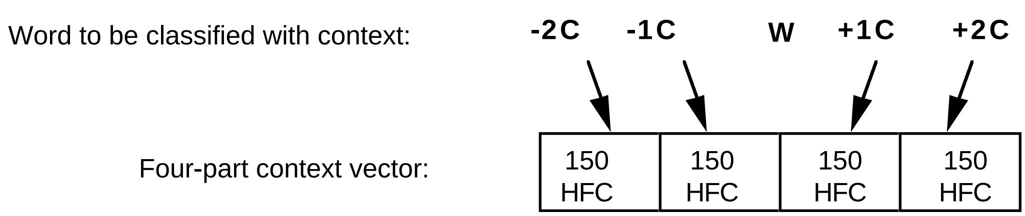

Figure 1. The 600-dimensional context vector around a target word W. Each subvector describes the frequency and mutual information of the occurrences of the 150 highest frequency words, HFC, in the corpus.

Subsequently, a 600-dimensional vector of mutual information values, *MI*, is computed from the frequencies as follows,

$$MI(cw) = \log_2\left(\frac{Nf_{cw}}{f_c f_w} + 1\right)$$

This expresses the mutual information value for the context word *c* appearing with the target word *w*. The mutual information is large whenever a context word appears at a much higher frequency, f_{cw}, in the neighborhood of a target word than would be predicted from the overall frequencies in the corpus, f_c and f_w. The formula adds 1 to the frequency ratio, so that a 0 (zero) occurrence corresponds to 0 mutual information. A possibly better strategy (Church, Gale, Hanks, & Hindle, 1991) is capable of generating negative mutual information for the non-occurrence or low-frequency occurrence of a very high-frequency word and has the form,

$$MI(cw) = \log_2\left(\frac{N(f_{cw} + 1)}{f_c f_w}\right)$$

In any case, some smoothing is necessary to prevent the mutual information from diverging when $f_{cw} = 0$.

SIMILARITY, CLUSTERING AND CLASSIFICATION IN WORD SPACE

When the mutual information vectors are computed for a number of words, they can be compared to see which words have similar contexts. The comparison we chose is the inner product, or cosine measure, which can vary between -1.0 and +1.0 (Myaeng & Li, 1992). Once this similarity is computed for all word pairs in a set, various techniques can be used to identify classes of similar words. The method we chose is hierarchical agglomerative clustering (Jain & Dubes, 1988). The two words with the highest similarity are first joined into a two-word cluster. A mutual information vector for the cluster is computed and the cluster and remaining words are again compared, choosing the most similar to join, and so on. (To compute the new mutual information vector, the context frequencies in the vectors for the two words or clusters joined at each step are summed, element-wise.) In this way, a binary tree is constructed with words at the leaves leading to a single root covering all words.

Each cluster, starting with individual target words has a frequency which is the sum of the frequencies of its two child clusters. Therefore, *every class generated is viewed extensionally*; a given class is not just a symbol, but rather a structured collection of occurrences in the corpus, with their attendant frequencies and contexts.

EXPERIMENT #1: CLASSIFICATION OF THE 1,000 HIGHEST FREQUENCY WORDS

The first experiment classified the 1,000 highest frequency words in the corpus, producing 999 clusters (0-998) during the process. $pre... and $pos... words were included in the context set, but not in the target set. Near the leaves, words clustered by syntax (part of speech) *and* by semantics. Later, larger nodes tended to contain words of the same syntactic class, but with less semantic homogeneity. *In each example below, the words listed are the entire contents of the node mentioned.* The most striking property of the clusters produced was the classification of words into coherent semantic fields. Grefenstette has pointed out (Grefenstette, 1992) that the *Deese antonyms*, such as "large" and "small" or "hot" and "cold" show up commonly in these analyses. Our methods discovered entire graded fields, rather than just pairs of opposites. The following example shows a cluster of seventeen adjectives describing comparative quantity terms, node 756, similarity 0.28,

decreased, effective, few, greater, high, higher, increased, large, less, low, lower, more, much, no, normal, reduced, short

Note that pairs such as "high" and "higher" and "low" and "lower" appear. "No", corresponding in meaning to "none" in this collection, is located at one extreme. The somewhat marginal item, "effective", entered the cluster late, at node 704. It appears in collocations such as "as effective as" and "effective than" which the other terms also appear in. Comparing the cluster to Roget's [Berrey, 1962 #868] we find that all the items are in the Roget category *Comparative Quantity* except for "effective" and "no". "Large" is not in this Roget category but Roget's lists "big", "huge" and "vast" there, so the omission is clearly an error in Roget's. With this correction, 88% (15/17) of the items are in a single Roget category.

The classification of technical terms from genetics and biochemistry is of particular interest, because many of these terms do not appear in available dictionaries or thesauri. Cluster 374, similarity 0.37 contains these 18 items,

che, cheA, cheB, cheR, cheY, cheZ, double, fla, flaA, flaB, flaE, H2, hag, mot, motB, tar, trg, tsr

All of these are abbreviations for specific bacterial mutations, except for "double". Its appearance drives home the point that the classification depends entirely on *usage*. 20 of the 30 occurences of "double" precede the words "mutant" or "mutants", as do most of the other mutation terms in this cluster.

Cluster 240, similarity 0.4 contains these terms,

microscopy, electrophoresis, chromatography

Each of these is a noun describing a common technique used in experiments in this domain.

The standard Linnean nomenclature of *Genus* followed by *species*, such as *Escherichia coli*, is reflected by cluster 414 that contains 22 species names and cluster 510 that contains 9 genus names .

In scientific research, the determination of causal factors and the discovery of essential elements is a major goal. Here are six concepts in this semantic field comprising node 183, similarity 0.43:

required, necessary, involved, responsible, essential, important

These terms are used almost interchangably in our corpus, but they don't fare as well in Roget's because of anthropocentric attachments to concepts such as fame, duty and legal liability.

Discussion of Experiment #1.

Given the limited context, the classification algorithm is bound to make mistakes, though a study of the text concordance will always tell us why the algorithm failed in any specific case. For example, as the similarity drops to 0.24 (node 824) we see the adverb triple "greatly", "rapidly", "almost", which is still acceptable, but by node 836 (similarity 0.24) we see the triple, "them", "ring", "rings". This pronoun and the two nouns appear to have been grouped together because they are often sentence-final or followed by commas. At the end there is only a single node, 998, which must include all words. It comes together stubbornly with a negative similarity of -0.51. One problem encountered in this work was that the later, larger clusters have much less coherence than we would hope for, identifying an important research issue. Experiment #1 took 20 hours to run on a Symbolics XL1200.

A fundamental problem is to devise decision procedures that will tell us which classes are semantically or syntactically homogeneous — where to cut the tree. The nice examples shown earlier broke down soon after, when words or clusters began to be added that in our judgement were weakly related. We are exploring the numerous methods to refine clusters once formed as well as to validate clusters for homogeneity (Jain & Dubes, 1988). There are also resampling methods to validate clusters formed by top-down methods (partitioning) (Jain & Moreau, 1987). All of these methods are computationally demanding but they can result in criteria for when to stop clustering. On the other hand, we mustn't assume that word relations are so simple that we can legitmately insist on finding neatly separated clusters. Word relations may simply be too complex and graded for this to ever occur.

It is clear to us why the semantic fields we discovered were not confined to synonyms. To understand this, consider the sentences, "The temperature is higher today." and, "The temperature is lower today." There is no way to tell from the syntax which word to expect. The choice is dependent on the situation in the world; it represents data from the world. The utterances are informative for just that reason. Information theory would suggest that for two contrasting words to be maximally informative, they should appear about equally often in discourse. This is born out in our corpus (f_{higher}=58, f_{lower}=46) and for the Brown corpus (f_{higher}=147, f_{lower}=110). The same relations are found for many other contrasting pairs, with some bias towards "positive" terms (the most extreme bias in our corpus is $f_{possible}$=88, $f_{impossible}$=0; "never say never" seems to be the catchphrase here — highly appropriate for the field of biology).

Some of the chemical term clusters that were generated are interesting because they contain terms such as "sugar" and "ion" along with specific members of the classes (hyponyms), such as "maltose" and "Na+" ("Na^+" when properly notated). Comparing these in our KWIC concordance suggests that there may be methodical techniques for identifying some of these generalization hierarchies by using machine learning (supervised classification) [Futrelle, 1993 #869]. For another discussion of attempts to generate generalization hierarchies, see (Myaeng & Li, 1992).

The hierarchical structures developed by our classification techniques, with each node characterized by a composite context vector, are similar to, and in fact can be used as, *decision trees*, which have been much studied in the machine learning literature (Quinlan, 1993). Given a new word, it can be successively compared to nodes in the tree, starting at the root, using a logarithmic search to find the best matching class. We do not have to make a prior commitment to labeled classes. The search discovers the optimal attachment for the new word, the most similar class.

EXPERIMENT #2: DISAMBIGUATION OF *-ED* FORMS

The following experiment is interesting because it shows a specific use for the similarity computations. They are used here to increase the accuracy of separating ambiguous terms into distinct classes, the tagging task. Again, this is a bootstrap method; no prior tagging is needed to construct the classes. But if we do identify the tags for a few items by hand or by using a hand-tagged reference corpus, the tags for all the other items in a cluster can be assumed equal to the known items.

The passive voice is used almost exclusively by the corpus, with some use of the editorial "We". This results in a profusion of participles such as "detected", "sequenced" and "identified". But such *-ed* forms can also be simple past tense forms or adjectives. In addition, we identified their use in a postmodifying participle clause such as, "... the value deduced from this measurement." Each one of the 88 instances of "cloned" and the 50 instances of "deduced" was given an instance label that identified its part of speech and unique identifying number and clustering was applied to the resulting collection, giving the result shown in Figure 2A. Experiments #2 and #3 took about 15 minutes each to run.

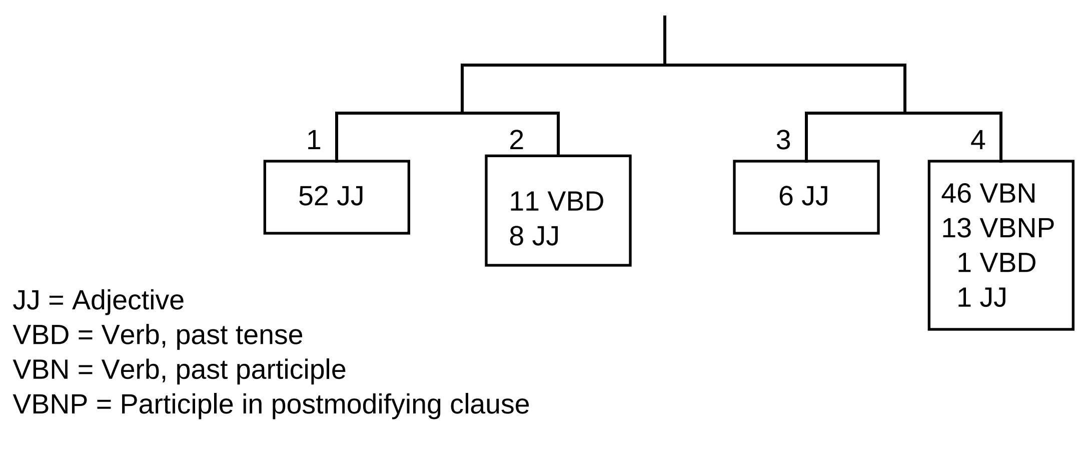

Figure 2A. Clustering of 88 occurrence of "cloned" and 50 occurrences of "deduced" into four syntactic categories. The abbreviations are based on (Francis & Kucera, 1982). There is a strong admixture of adjectives in node 2 and all the postmodifiers are confounded with the past participles in node 4. The total number of errors (minority classes in a node) is 23 for a success rate of (138-23) / 138 = 83%.

The results shown in Figure 2A can be improved as follows. Because we are dealing with single occurrences, only one element, at most, in each of the four context word vectors is filled, with frequency 1, the other 149 elements have frequency (and mutual information) 0.0. These sparse vectors will therefore have little or no overlap with vectors from other occurrences. In order to try to improve the classification, we expanded the context values in an effort to produce more overlap, using the following strategy: We proceed as if the corpus is far larger and in addition to the actual context words already seen, there are many occurrences of highly similar words in the same positions. So for each non-zero context in each set of 150, we expand it to an ordered class of similar words in the 150, picking words above a fixed similarity threshold (0.3 for the experiments reported here). Such a class is called a *simset*, made up of a *base word* and a sequence of *expansion words*.

The apparent frequency of each expansion word is based on its corpus frequency relative to the corpus frequency of the word being expanded. To expand a single context word instance c_i appearing with frequency f_{ik} in the context of 1 or more occurrences of center word w_k, choose all c_j such that $c_j \in$ {set of high-frequency context words} and the similarity $S(c_i,c_j) \geq S_t$, a threshold value. Set the apparent frequency of each expansion word c_j to $f_{jk} = S(c_i,c_j) * f_{ik} * f_j / f_i$, where f_i and f_j are the corpus frequencies of c_i and c_j. Normalize the total frequency of the context word plus the apparent frequencies of the expansion words to f_{ik}. For the example being discussed here, $f_{ik} = 1$, $S_t = 0.3$ and the average number of expansion words was 6.

Recomputing the classification of the *-ed* forms with the expanded context words results in the improved classification shown in Figure 2B. The number of classification errors is halved.

Discussion of Experiment #2.

This analysis is very similar to part-of-speech tagging. The simsets of only 6 items are far smaller than the part-of-speech categories conventionally used. But since we use high frequency words, they represent a substantial portion of the instances. Also, they have higher specificity than, say, *Verb*. Many taggers work sequentially and depend on the left context. But some words are best classified by their left context and some by their right — we supply both. Clearly this small experiment did not reach the accuracy of the very best taggers, but it performed well.

This experiment has major ramifications for the future. The initial classifications we did merged all identical word forms together, both as targets and contexts. But disambiguation techniques such as in Experiment #2 can be used to differentially tag word occurrences, with some degree of accuracy. These newly classified items can in turn be used as new target and context items (if their frequencies are adequate) and the analysis can be repeated. Iterating the method in this way should be able to refine the classes until a fixed point is reached at which no further changes in the classification occur. The major challenge in using this approach will be to keep it computationally tractable. This approach is similar in spirit to the iterative computational approaches of the hidden Markov models (Kupiec, 1989; Kupiec, 1992; Rabiner, 1989), though our zeroth order solution begins quite close to the desired result, so it should converge very close to a global optimum.

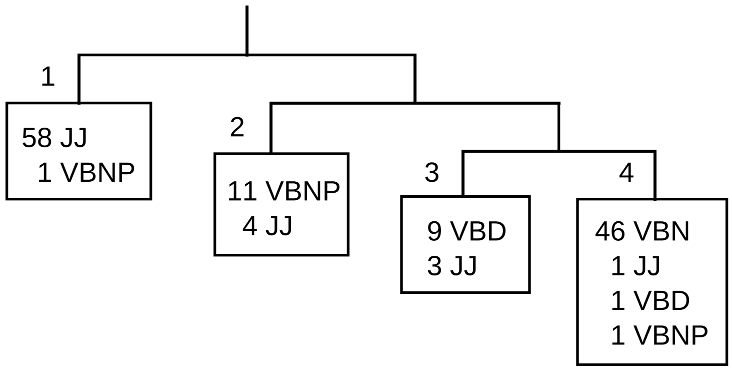

Figure 2B. Clustering of "cloned" and "deduced" after expansion of the context words. The postmodifying form, not isolated before, is fairly well isolated in its own subclass. The total number of errors is reduced from 23 to 11, for a success rate of 92%.

EXPERIMENT #3: CLASSIFICATION OF SINGLE WORD OCCURRENCES

When classifying multiple instances of a single word form as we did in the previous example, there are numerous collocations that aid the classification. For example, there are 16 occurrences in the corpus of the phrase, "of the deduced amino acid sequence". But with words of frequency 1, we cannot rely on such similarities. Nevertheless, we experimented with classifying 100 words of corpus frequency 1 with and without expanding the context words. Though hand scoring the results is difficult, we estimate that there were 8 reasonable pairs found initially and 26 pairs when expansion was used.

Examples of words that paired well without expansion are "overlaps" and "flank" (due to a preceding "which") and "malB" and "cheA-cheB" (due to the context "...the [malB,cheA-cheB] region..."). After expansion, pairs such as "setting", "resetting" appeared (due in part to the expansion of the preceding "as" and "to" context words into sets which both included "with", "in" and "by").

Discussion of Experiment #3.

The amount of information available about frequency 1 word can vary from a lot to nothing at all, and most frequently tends to the latter, viz., "John and Mary looked at the blork." Nevertheless, such words are prominent, 44% of our corpus' vocabulary. About half of them are non-technical and can therefore be analyzed from other corpora or on-line dictionaries. Word morphology and latinate morphology in particular, will chip away further. Online chemical databases, supplemented with rules for chemical nomenclature will clarify additional items, e.g., "2-epoxypropylphosphonic" or "phosphoglucomutase-deficient". There are naming conventions for genetic strains and mutants which aid recognition. So this class of words should yield to a war of attrition.

FURTHER DISCUSSION AND FUTURE DIRECTIONS

Our corpus lies at one end of the spectrum compared to ones of 40 million words (Finch & Chater, 1992) or even 360 million (Brown, Della Pietra, deSousa, Lai, & Mercer, 1992). Judging by the results we have presented, especially for the full 1,000 word clustering, our corpus appears to make up in specificity for what it lacks in size. Extending this work beyond abstracts to full papers will be challenging because our corpus requires SGML markup to deal with Greek characters, superscripts and subscripts, etc. (Futrelle, Dunn, Ellis, & Pescitelli, 1991). We have over 500,000 words from the bacterial chemotaxis research papers carefully marked up by hand in this way.

The characterization of context can obviously be extended to more context positions or words, and extensions of our word-rooted expansion techniques are potentially very powerful, combining broad coverage with specificity in a "tunable" way. Morphology can be added to the context vectors by using the ingenious suggestion of Brill to collect high-frequency tri-letter word endings (Brill & Marcus, 1992).

One of the more subtle problems of the context specification is that it uses summed frequencies, so it may fail to retain important collocations. Thus if only AB or CD sequences occurred or only AD or CB sequences occurred, they would lead to the same (summed) context vector. The only correlations faithfully retained are those with the target word. Characterizing context n-grams could help work around this problem, but is a non-trivial task.

Clearly, the methods described here will be able to make a substantial contribution to the difficult task of extracting lexical knowledge from text.

BIBLIOGRAPHY

Brill, E., & Marcus, M. (1992). Tagging an Unfamiliar Text with Minimal Human Supervision. In AAAI Fall Symposium Series: Probabilistic Approaches to Natural Language (Working Notes), (pp. 10-16). Cambridge, MA:

Brown, P. F., Della Pietra, V. J., deSousa, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-Based *n*-gram Models of Natural Language. Computational Linguistics, 18(4), 467-479.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Eds.), *Lexical Acquisition:* Exploiting On-Line Resources to Build a Lexicon (pp. 115-164). Hillsdale, NJ: Lawrence Erlbaum.

Church, K. W. (1988). A Stochastic Parts Program and Noun Parser for Unrestricted Text. In Proc. 2nd Conf. on Applied Nat. Lang. Processing, (pp. 136-143). Austin, TX:

Church, K. W., & Hanks, P. (1990). Word Association Norms, Mutual Information and Lexicography. Computational Linguistics, 16(1), 22-29.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A Practical Part-of-Speech Tagger. In Proc. 3rd Conf. on Applied Natural Language Processing, (pp. 133-140).

Finch, S., & Chater, N. (1992). Bootstrapping Syntactic Categories Using Statistical Methods. In W. Daelemans & D. Powers (Ed.), Proc. 1st SHOE Workshop, (pp. 229-235). Tilburg U., The Netherlands:

Francis, W. N., & Kucera, H. (1982). Frequency Analysis of English Usage. Boston, MA: Houghton Mifflin.

Futrelle, R. P., Dunn, C. C., Ellis, D. S., & Pescitelli, M. J., Jr. (1991). Preprocessing and lexicon design for parsing technical text. In Proc. 2nd Intern'l Workshop on Parsing Technologies, (pp. 31-40). ACL.

Futrelle, R. P., Kakadiaris, I. A., Alexander, J., Carriero, C. M., Nikolakis, N., & Futrelle, J. M. (1992). Understanding Diagrams in Technical Documents. IEEE Computer, 25(7), 75-78.

Gale, W. A., Church, K. W., & Yarowsky, D. (1992). Work on Statistical Methods for Word Sense Disambiguation. In AAAI Fall Symposium Series: Probabilistic Approaches to Natural Language (Working Notes), (pp. 54-60). Cambridge, MA:

Gauch, S., & Futrelle, R. P. (1993). The Scientist's Assistant: Structure-Based Navigation of Technical Texts. In 2nd Ann. Symp. on Document Analysis and Information Retrieval, . Las Vegas, NV: IEEE (pp. 367-378).

Grefenstette, G. (1992). Finding Semantic Similarity in Raw Text: the Deese Antonyms. In AAAI Fall Symposium Series: Probabilistic Approaches to Natural Language (Working Notes), (pp. 61-66). Cambridge, MA:

Jain, A. K., & Dubes, R. C. (1988). Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice Hall.

Jain, A. K., & Moreau, J. V. (1987). Bootstrap Technique in Cluster Analysis. Pattern Recognition, 20(5), 547-568.

Krovetz, R. (1991). Lexical Acquisition and Information Retrieval. In U. Zernik (Eds.), Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon (pp. 45-64). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Kupiec, J. (1989). Augmenting a hidden Markov model for phrase dependent word tagging. In DARPA Speech and Natural Language Workshop, a (pp. 92-98). Cape Cod MA:

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. Computer Speech and Language, 6, 225-242.

Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from and Ice Cream Cone. In ACM SIGDOC Conf., (pp. 24-26). Toronto, Ont.: ACM Press.

Liddy, E. D., & Paik, W. (1992). Statistically-Guided Word Sense Disambiguation. In AAAI Fall Symposium Series: Probabilistic Approaches to Natural Language (Working Notes), (pp. 98-107). Cambridge, MA:

Myaeng, S. H., & Li, M. (1992). Building Term Clusters by Acquiring Lexical Semantics from a Corpus. In Y. Yesha (Ed.), CIKM-92, (pp. 130-137). Baltimore, MD: ISMM.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models an Selected Applications in Speech Recognition. Proceedings of the IEEE, 77(2), 257-286.

Schutze, H. (1992). Context Space. In AAAI Fall Symposium Series: Probabilistic Approaches to Natural Language (Working Notes), (pp. 113-120). Cambridge, MA:

Zernik, U. (1991). Train1 vs. Train2: Tagging Word Senses in Corpus. In U. Zernik (Eds.), Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon (pp. 97-112). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.