Discrimination Decisions for 100,000-Dimensional Spaces

William A. Gale
Kenneth W. Church
David Yarowsky

AT&T Bell Laboratories

e-mail: {*gale, kwc, yarowsky*}@*research.att.com*

Abstract

Discrimination decisions arise in many natural language processing tasks. Three classical tasks are discriminating texts by their authors (author identification), discriminating documents by their relevance to some query (information retrieval), and discriminating multi-meaning words by their meanings (sense discrimination). Many other discrimination tasks arise regularly, such as determining whether a particular proper noun represents a person or a place, or whether a given word from some teletype text would be capitalized if both cases had been used.

We (1992) introduced a method designed for the sense discrimination problem. Here we show that this same method is useful in each of the five text discrimination problems mentioned.

We also discuss areas for research based on observed shortcomings of the method. In particular, an example in the author identification task shows the need for a robust version of the method. Also, the method makes an assumption of independence which is demonstrably false, yet there has been no careful study of the results of this assumption.

1 Introduction

Statistical methods are being applied to more and more problems in natural language. Although there has been a long tradition of statistical work in natural language (e.g., Zipf 1932, Yule 1944, Mosteller and Wallace 1964, Salton and Yang 1973, Harris 1968), there has recently been a revival of interest in statistical approaches to natural language because computer readable text is becoming easier to obtain in large quantities. Just ten years ago, the one million word Brown Corpus (Francis and Kučera, 1982) was considered large. These days, a corpus has to be at least ten times larger in order to be considered large. And some researchers are using corpora that are a hundred times larger than the Brown Corpus.

There are a number of well-known applications of discrimination techniques in natural language processing, especially information retrieval and author identification. A number of other problems can be addressed with very similar discrimination techniques, especially the very important problem of sense disambiguation. We have also applied similar discrimination techniques to restore capitalization in upper-case only text and to distinguish names of places from names of people.

It is an interesting question whether techniques that were developed several decades ago will continue to scale up as we continue to look at larger and larger problems. One might wonder if it is appropriate to expect a discrimination technique that was developed on a "small" problem such as the Federalist Papers to work on a "large" corpus of some tens or hundreds of millions of words of text. Most of these discrimination techniques use a V-dimensional space, where V is size of the vocabulary. The vocabulary and the dimensionality of the problem grow rapidly as we consider larger and larger corpora. The old Federalist collection contains a mere $N = 208,304$ tokens and a vocabulary of only $V = 8663$ types; more modern corpora contain some tens of millions of tokens and hundreds of thousands of types.

Therefore we must find discrimination techniques for dealing with spaces having about 100,000 dimensions. These methods can either be "direct," not reducing the number of dimensions of the space, or "indirect," reducing the number of dimensions to some manageable number. We find that Bayesian decision analysis can be used in a direct fashion for each of the problems we examine. Indirect approaches have been necessary due to computing constraints until recently, so there is some heuristic experience with them. However, principled study of indirect approaches has been possible only since direct approaches could be implemented, which is to say recently, so little is known about them.

2 Discrimination Problems in Natural Language Processing

Text discrimination problems begin by specifying a corpus, a collection of documents such as the Federalist Papers, newswire stories collected from Associated Press (AP) over a few years, the official record of the Canadian parliamentary debates, or a set of encyclopedia articles. Documents are represented as a sequence of tokens, e.g., words, punctuation, type-setting marks, and delimiters to mark sentences and paragraphs.

In the training phase, we are given two (or more) sets of documents and are asked to construct a discriminator which can distinguish between the two (or more) classes of documents. These discriminators are then applied to new documents during the testing phase. In the author identification task, for example, the training set consists of several documents written by each of the two (or more) authors. The resulting discriminator is then tested on documents whose authorship is disputed. In the information retrieval application, a query is given and the training set consists of a set of one or more documents relevant to the query and a set of zero or more irrelevant documents. The resulting discriminator is then applied to all documents in the library in order to separate the more relevant ones from the less relevant ones. In the sense disambiguation case, the 100-word context surrounding instances of a polysemous word (e.g., *bank*) can be treated very much like a document, as we will see.

There is an embarrassing wealth of information in the collection of documents that could be used as the basis for discrimination. To date, most researchers using statistical techniques have tended to treat documents as "merely" a bag of words, and have generally tended to ignore much of the linguistic structure, especially dependencies on word order and correlations between pairs of words. The collection of documents can then be represented as a term by document matrix, where each cell counts the number of times that a particular term appears in a particular document. Since there are $V \approx 100,000$ terms, the term by document matrix contains a huge amount of information, even allowing for the fact that the matrix is quite sparse and many of the cells are empty.

One approach to these problems has been Bayesian discrimination analysis. Mosteller and Wallace (1964, section 3.1) used the following formula to combine new evidence (e.g., the term by document matrix) with prior evidence (e.g., the historical record) in their classic authorship study of the Federalist Papers.

$$final\ odds = (initial\ odds) \times (likelihood\ ratio)$$

For two groups of documents, the equation becomes

$$\frac{P(class_1)}{P(class_2)} = \frac{p(class_1)}{p(class_2)} \times \frac{L(class_1)}{L(class_2)}$$

where P represents a final probability, p represents an initial probability, and L represents a likelihood. Similar equations can be found in textbooks on information retrieval (e.g., Salton 1989, equation 10.17).

The initial odds depend on the problem. In the author identification problem, for example, the initial odds are used to model what we know about the documents from the various conflicting historical records. In the information retrieval application, the user may have a guess about the fraction of the library that he or she would expect to be relevant; such a guess could be used as the prior. The objectivity of the prior depends on the problem. In the author identification case, it is largely subjective. For the information retrieval problem, a baseline probability could be established from past experience in the number of relevant documents found. For many other problems, including spelling correction, sense disambiguation, and other problems discussed here, the prior can be quite objective and very useful.

It is common practice to decompose the likelihoods into a product of likelihoods over tokens in the document (under appropriate independence assumptions):

$$\frac{L(class_1)}{L(class_2)} \approx \prod_{token\ \text{in}\ doc} \frac{Pr(token|class_1)}{Pr(token|class_2)}$$

The crucial ingredients for this calculation are the probabilities of each term in the vocabulary *conditional* on the document being from a given class. These conditional probabilities have been computed in a number of different ways depending on the application and the study. In the next section we will introduce a novel method of calculating these conditional probabilities. The method was originally designed for the sense disambiguation application, though we have found that the method can be used "off-the-shelf" to produce results that are comparable to (though perhaps not quite as good as) methods that have been highly tuned over many years for a particular problem.

3 Sense Discrimination: An Example of the Approach

Consider, for example, the word *duty*, which has at least two quite distinct senses: (1) a tax and (2) an obligation. Three examples of each sense are given below in Table 1.

Table 1: Sample Concordances of *duty* (split into two senses)

Sense	Examples (from Canadian Hansards)
tax	companies paying **duty** and then claiming a refund
	a countervailing **duty** of 29,1 per cent on canadian
	states imposed a **duty** on canadian saltfish last year
obligation	is my honour and **duty** to present a petition duly
	beyond the call of **duty** ? SENT i know what time
	addition , it is my **duty** to present the government 's

Sense disambiguation has been recognized as a major problem in natural language processing research for over forty years. There has been a considerable body of work on the subject, but much of the work has been stymied by difficulties in acquiring appropriate lexical resources (e.g., semantic networks, annotated corpora, dictionaries, thesauruses, etc.). In particular, much of the work on qualitative methods has had to focus on "toy" domains since currently available semantic networks generally lack the broad coverage that would be required to address a realistic problem. Similarly, much of the work on quantitative methods has had to depend on small amounts of hand-labeled text for testing and training.

We achieved considerable progress as reported in (Gale et al., 1992) by taking advantage of a new source of testing and training materials for studying sense disambiguation methods. Rather than depend on small amounts of hand-labeled text, we used relatively large amounts of parallel text, text such as the Canadian Hansards, which are available in multiple languages. The translation can often be used in lieu of hand-labeling. For example, the two senses of *duty* mentioned above are usually translated with different French words in the French version. The tax sense of *duty* is typically translated as *droit* whereas the obligation sense is typically translated as *devoir*. Thus, we can collect a number of tax sense instances of *duty* by extracting instances of *duty* that are translated with *droit*, and we can collect a number of obligation instances by extracting instances that are translated with *devoir*. In this way, we were able to acquire considerable amounts of testing and training material for study of quantitative methods. More details on the preparation of the testing and training materials can be found in (Gale and Church, 1991a, 1991b).

The availability of this testing and training material has enabled us to develop quantitative disambiguation methods that achieve 92 percent accuracy in discriminating between two very distinct senses of a noun such as *duty*. While the 8% error rate appears to be about half that of disambiguation methods published before we bagan the work, it is even more important that the proposed method has a better potential of scaling up to handle realistic size vocabularies of tens of thousands of ambiguous words. There have been several other studies of sense disambiguation recently (Brown et al., 1991), (Dagan, Itai, and Schwall, 1991), (Hearst, 1991), (Zernik, 1992), (Yarowsky, 1992), and (Leacock et al., 1993).

We made a number of studies, most of which focus on the six English nouns shown in Table 2 (below). This table also shows the two French translations and an English gloss of the relevant sense distinction.

Table 2: Six Polysemous Words

English	French	sense	N	% correct
duty	droit	tax	1114	97
	devoir	obligation	691	84
drug	médicament	medical	2992	84
	drogue	illicit	855	97
land	terre	property	1022	86
	pays	country	386	89
language	langue	medium	3710	90
	langage	style	170	91
position	position	place	5177	82
	poste	job	577	86
sentence	peine	judicial	296	97
	phrase	grammatical	148	100
Average				90
Average	(with prior)			92

For two senses, the Bayesian equation mentioned above becomes:

$$\frac{P(sense_1)}{P(sense_2)} = \frac{p(sense_1)}{p(sense_2)} \times \frac{L(sense_1)}{L(sense_2)}$$

where p, P and L are the initial probability, the final probability and likelihood, as before. The initial probabilities are determined from the overall frequencies of the two senses in the corpus. As in other large dimension discrimination problems, the likelihoods are decomposed into a product over tokens:

$$\frac{L(sense_1)}{L(sense_2)} \prod_{token \text{ in } context} \frac{Pr(token|sense_1)}{Pr(token|sense_2)}$$

As mentioned above, this model ignores a number of important linguistic factors such as word order and collocations (correlations among words in the context). Nevertheless, there are $2V \approx 200,000$ parameters in the model. It is a non-trivial task to estimate such a large number of parameters, especially given the sparseness of the training data. The training material typically consists of approximately 12,000 words of text (100 words of context for 60 instances of each of two senses). Thus, there are more than 15 parameters to be estimated for each data point. Clearly, we need to be fairly careful given that we have so many parameters and so little evidence.

The conditional probabilities, $Pr(token|sense)$, can be estimated in principle by selecting those parts of the entire corpus which satisfy the required conditions (e.g., 100-word contexts surrounding instances of one sense of *duty*), counting the frequency of each word, and dividing the counts by the total number of words satisfying the conditions. However, this estimate, which is known as the maximum likelihood estimate (MLE), has a number of well-known problems. In particular, it will assign zero probability to words that do not happen to appear in the sample. Zero is not only a biased estimate of their true probability, but it is also unusable for the sense disambiguation task (and for quite a number of other applications). In addition, MLE also produces poor estimates for words that appear only once or twice in the sample. In another application (spelling correction), we have found that poor estimates of context are worse than none; that is, at least in this application, we found that it would be better to ignore the context than to model it badly with something like MLE (Gale and Church, 1990).

The method derived in the next section was introduced by Gale, Church and Yarowsky (1992) and uses the information from the entire corpus in addition to information from the conditional sample in order to avoid these problems. We will estimate $Pr(token|sense)$ by interpolating between word probabilities computed within the 100-word context and word probabilities computed over the entire corpus. For a word that appears fairly often within the 100-word context, we will tend to believe the local estimate and will not weight the global context very much in the interpolation. Conversely, for a word that does not appear very often in the local context, we will be much less confident in the local estimate and will tend to weight the global estimate somewhat more heavily. The key observation behind the method is this: the entire corpus provides a set of well measured probabilities which are of unknown relevance to the desired conditional probabilities, while the conditional set provides poor estimates of probabilities that are certainly relevant. Using probabilities from the entire corpus thus introduces bias, while using those from the conditional set introduce random errors. We seek to determine the relevance of the larger corpus to the conditional sample in order to make this trade off between bias and random error.

The interpolation procedure makes use of a prior expectation of how much we expect the local probabilities to differ from the global probabilities. Mosteller and Wallace "expect[ed] both authors to have nearly identical rates for almost any word" (p. 61). In fact, just as they had anticipated, we have found that only 2% of the vocabulary in the Federalist corpus has significantly (3 standard deviation) different probabilities depending on the author. Moreover, the most important words for the purposes of author identification appear to be high frequency function words. Our calculations show that *upon, of* and *to* are strong indicators for Hamilton and that *the, and, government* and *on* are strong indicators for Madison. These are all high frequency function words (at least in these texts), with the exception of *government*, which is, nonetheless, extremely common and nearly devoid of content.

In contrast, we expect fairly large differences in the sense disambiguation application. For example, we find that the tax sense of *duty* tends to appear near one set of content words (e.g., *trade* and *lumber*) and that the obligation sense of *duty* tends to appear near quite a different set of content words (e.g., *honour* and *order*), at least in the Hansard corpus. Approximately 20% of the vocabulary in the Hansards has significantly different probabilities near *duty* than otherwise. In short, the prior expectation depends very much on the application. In any particular application, we set the prior by estimating the fraction of the vocabulary whose conditioned probabilities differ significantly from the global probabilities. Thus the same interpolation procedure is used for all of the applications discussed here.

4 The Interpolation Procedure

Let the entire corpus be divided into a sample, satisfying some condition, of size n, and the residual corpus (the entire corpus less the conditional sample) of size $N \gg n$. Let a be the frequency of a given word in the conditional sample, and A its frequency in the residual corpus. Either of these frequencies may be zero, but not both. Let π represent the probability of the word given the condition establishing the sample. Before knowing the frequency of the word in either the sample or the residual corpus, we could express our ignorance of the value of π by the *uninformative distribution*:

$$B^{-1}(\frac{1}{2},\frac{1}{2})\pi^{-\frac{1}{2}}(1-\pi)^{-\frac{1}{2}}$$

where $B(x,y)$ is the Beta function. Several variations of the method can be based on variations in the uninformative distribution. If A instances out of N independent observations relevant to the determination of π were found, then the distribution expressing our knowledge would become

$$B^{-1}(A+\frac{1}{2},N-A+\frac{1}{2})\pi^{A-\frac{1}{2}}(1-\pi)^{N-A-\frac{1}{2}}$$

While we have A out of N observations of the word in question in the residual corpus, we do not know their relevance. Thus we set as our knowledge before observing the conditional sample the distribution:

$$\begin{aligned} p(\pi) &= rB^{-1}(A+\frac{1}{2},N-A+\frac{1}{2})\pi^{A-\frac{1}{2}}(1-\pi)^{N-A-\frac{1}{2}} \\ &\quad + (1-r)B^{-1}(\frac{1}{2},\frac{1}{2})\pi^{-\frac{1}{2}}(1-\pi)^{-\frac{1}{2}} \end{aligned}$$

where $0 \leq r \leq 1$ is interpreted as the relevance of the residual corpus to the conditional sample. When $r = 0$, this gives the uninformative distribution, while if $r = 1$, it gives the distribution after observing the residual corpus. Another way of reading this is that with probability r we are expecting an observation in line with the residual corpus, but that with probability $1 - r$ we won't be surprised by any value.

The joint probability of observing a out of n instances of the word in question in the conditional sample and that the conditional probability is π is then

$$p(\pi,a)=\binom{n}{a}\Big\{rB^{-1}(A+\frac{1}{2},N-A+\frac{1}{2})\pi^{A+a-\frac{1}{2}}(1-\pi)^{N-A+n-a-\frac{1}{2}}$$

$$+(1-r)B^{-1}(\frac{1}{2},\frac{1}{2})\pi^{a-\frac{1}{2}}(1-\pi)^{n-a-\frac{1}{2}}\Big\}$$

We then form

$$p(a) = \int_0^1 p(\pi, a) d\pi$$

and

$$p(\pi|a) = \frac{p(\pi, a)}{p(a)}$$

which is then integrated to give

$$E(\pi|a) = \int_0^1 \pi p(\pi|a) d\pi = \frac{r\frac{B(A+a+1\frac{1}{2},N-A+n-a+\frac{1}{2})}{B(A+\frac{1}{2},N-A+\frac{1}{2})} + (1-r)\frac{B(a+1\frac{1}{2},n-a+\frac{1}{2})}{B(\frac{1}{2},\frac{1}{2})}}{r\frac{B(A+a+\frac{1}{2},N-A+n-a+\frac{1}{2})}{B(A+\frac{1}{2},N-A+\frac{1}{2})} + (1-r)\frac{B(a+\frac{1}{2},n-a+\frac{1}{2})}{B(\frac{1}{2},\frac{1}{2})}}$$

This can be approximated in various ways, but it is practical to calculate it directly using the relationship

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

The parameter r, which denotes the relevance of the residual corpus to the conditional sample, can be estimated in various ways. Its basic interpretation is the fraction of words that have conditional probabilities close to their global probabilities (as estimated from the residual sample). Thus given a set of estimates of conditional probabilities, one can estimate r as the fraction of them which lie within a few standard deviations of the corresponding global probabilities. This estimate is performed using the words which are observed in the conditional sample. Alternatively r can be regarded as a free parameter of the method and adjusted to produce optimal results on a specific task. Although it could be varied for each word, we have used $r = 0.8$ for all words in the sense disambiguation application, and $r = 0.98$ for all words in the author identification application, based on empirical findings from the data.

5 Example of the Interpolation Procedure

Table 3 gives a sense of what the interpolation procedure does for some of the words that play an important role in disambiguating between the two senses of *duty* in the Canadian Hansards. Recall that the interpolation procedure requires a conditional sample. In this example, the conditional samples are obtained by extracting a 100-word window surrounding each of 60 training examples. The training examples were selected by randomly sampling instances of *duty* in the Hansards until 60 instances were found that were translated as *droit* and 60 instances were found that were translated as *devoir*. The first set of 60 are used to construct the model for the tax sense of *duty* and the second set of 60 are used to construct the model for the obligation sense of *duty*.

The column labeled "freq" shows the number of times that each word appeared in the conditional sample. For example, the count of 50 for the word *countervailing* indicates

Table 3: Selected Portions of Two Models

tax sense of *duty*			
weight*freq	weight	freq	word
285	5.7	50	countervailing
111.8	4.3	26	duties
99.9	2.7	37	u.s
73.1	1.7	43	trade
70.2	1.8	39	states
69.3	3.3	21	duty
68.4	3.6	19	softwood
68.4	1.9	36	united
58.8	8.4	7	rescinds
54	3.0	18	lumber
50.4	4.2	12	shingles
50.4	4.2	12	shakes
46.2	2.1	22	against
41.8	1.1	38	canadian

obligation sense of *duty*			
weight*freq	weight	freq	word
64	3.2	20	petitions
59.28	0.3	228	to
51	3.0	17	petition
47.6	2.8	17	pursuant
46.28	0.5	89	mr
37.8	2.7	14	honour
37.8	1.4	27	order
36	2.0	18	present
33.6	2.8	12	proceedings
31.5	3.5	9	prescription
31.32	0.9	36	house
29.7	3.3	9	reject
29.4	4.2	7	boundaries
28.7	4.1	7	electoral

that *countervailing* appeared 50 times within the 100-word window of an instance of *duty* that was translated as *droit*. This is a remarkable fact, given that *countervailing* is a fairly unusual word.

The second column (labeled "weight") models the fact that 50 instances of *countervailing* are more surprising than 228 instances of *to* (even though 228 instances of *to* is somewhat surprising). The weights for a word are its log likelihood in the conditional sample compared with its log likelihood in the global corpus. The first column, the product of these log likelihoods and the frequencies, is a measure of the importance, in the training set, of the word for determining which sense the training examples belong to. Note that words with large scores do seem to intuitively distinguish the two senses, at least in the Canadian Hansards. The set of words listed in Table 3 under the obligation sense of *duty* is reasonable given that the Hansards contain a fair amount of boilerplate of the form: "Mr. speaker, pursuant to standing order..., I have the honour and duty to present petitions duly signed by... of my electors...".

By interpolating between the local and global probabilities in this way, we are able to estimate considerably more parameters than there are data points (words) in the training corpus. The interpolation procedure assumes that one selection of natural language is roughly similar to another. In this way, it becomes feasible to estimate the $2V \approx 200,000$ parameters, one for each word in the vocabulary and each of the two senses. This is the basis for a direct approach to the problem of a high dimensional space using Bayesian decision analysis. We are, of course, assuming that conditional on being in the training sample, correlations between words are zero. Although this is a common assumption in information retrieval and author identification applications, it might be a cause of some concern. Fortunately, there are some reasons to believe that the correlations do not have a very large effect, which we will review in in Section 11. But first, let us describe some other applications of discrimination methods in language. We will start with the two very well established applications, author identification and information retrieval, and then we will move on to describe some applications that we have been working on in our lab.

6 Author Identification

The approach we have taken to discrimination in a high dimensional space was inspired, as we have said, by the classic work in author discrimination by Mosteller and Wallace (1964). After completing the study described in the previous section, we decided to spend a little time (approximately one day) investigating how well the same methods would work on Mosteller and Wallace's application. Although it may not be fair to compare a single day of work with a decade of research, we are excited by the fact that we could use the basic techniques "off-the-shelf" to produce nearly as good results, without spending any time tuning the methods to the particular application.

Mosteller and Wallace studied the Federalist papers, a collection of 85 documents debating the adoption of a new constitution by the American states. Of these, 51 are known to be by Hamilton (H), 14 are known to be by Madison (M), 5 by Jay (J), and 3 jointly by Madison and Hamilton (MH). The remaining 12 are of unknown authorship, but are presumed to be by either Hamilton or Madison (M?). Mosteller and Wallace found that all twelve would be ascribed to Madison if one's prior was indifferent between Hamilton and Madison before considering this evidence.

In the previous section, we set r, the fraction of words whose conditional probabilities are similar to the global probabilities, to 0.8, based on the observation that about 20% of the vocabulary has significantly different probabilities in the conditional sample than they would otherwise have. As mentioned above, we believe that r depends fairly strongly on the application, and after a quick investigation, we decided to set r to 0.98, based on the observation that only 2% of the Federalist vocabulary has significantly different probabilities depending on the author. Thus for this problem, r is determined from the data and is not subjective.

We then built a Hamilton model from all the known Hamilton papers, and a Madison model from all the known Madison papers. These models were applied to the remaining papers, namely J + MH + M?. A positive score would indicate authorship by Madison. All of the disputed papers (au = M?) scored positively. Thus, we reach the same conclusion as Mosteller and Wallace did. That is, assuming equal priors for the two papers, then we conclude that all of the disputed papers were more likely to have been written by Madison than by Hamilton after taking the word frequency evidence into account.

We also wanted to test the H and M papers. For each H paper, we built a model from all the Hamilton papers excepting the one in question, comparing the results to the overall Madison model. Likewise for each Madison paper we built such a cross validating model and compared results to the overall Hamilton model. Thereby, the scores for the known papers can be used to cross-check the validity of the method. Except for one mistake (document 47), all of the H papers scored negatively and all of the M papers scored positively, as they should.

Our cross-check is probably somewhat more thorough than the one that Mosteller and Wallace were able to perform forty years ago, since we now have the computer power to use a jackknifing technique and exclude each of the H and M papers from the models while they are being scored. If we had not taken this extra precaution, we would not have noticed the problem with document 47.

The one clear mistake is instructive. The problem with document 47 can be attributed to the single word *court*. The word *court* is mentioned 8 times in document 47, and in fact, document 47 is the only M document to mention the word *court* at all. And to make matters worse, *court* happens to be a favorite H word (Hamilton used it 90 times). If we remove that word, then we would find that the paper is very strongly in M's direction. Removing *court* has the effect of adding 68 to the score. Thus, instead of scoring -16 which is well within the H range, it would receive a strong M score of 42. This example suggests that we need to investigate robust approaches which would reduce the danger that one outlier could dominate the result.

7 Information Retrieval: Alternate Approaches

There have been quite a number of attempts to apply statistical techniques to problems in information retrieval. The *vector-space* model (Salton, 1989, section 10.1) is perhaps the most popular statistical method among researchers in information retrieval. Let q be a query, and let $D = d_i$ be a set of documents. The information retrieval task is to sort the documents by how similar they are to the query. According to the vector-space model, the query and the documents are represented as vectors. Each vector has V dimensions, where V is the vocabulary of terms. Thus, if the term, *strong*, for example, appears 5 times in a document, then its vector will have a magnitude of 5 along the dimension corresponding to *strong*. It is then a straightforward matter to sort vectors by a standard similarity measure such as cosine distance:

$$sim(x, y) = \frac{\sum_{i=1}^{V} x_i y_i}{|x|\,|y|}$$

where

$$|x| \;=\; \sqrt{\sum_{i=1}^{V} x_i^2}$$

The probabilistic retrieval model, (Salton, section 10.3), a rival to the vector-space model, sorts documents by how likely the words in the document were to have come from relevant documents $Pr(token|rel)$, as opposed to irrelevant documents $Pr(token|irrel)$. The probabilistic retrieval model shares many of the same fundamentals as the methods we have been studying, although the estimation of conditional probabilities is somewhat different in detail.

$$score(d_i) = \prod_{token \text{ in } d_i} \frac{Pr(token|rel)}{Pr(token|irrel)}$$

In our experiments, vector-space methods worked slightly better for information retrieval than our probability measure. (So naturally we tried them on the sense discrimination problem, but found them markedly inferior.) One factor that differentiates the problems is the length of query and of document. In the sense disambiguation task the query may consist of a hundred words of relevant context, while for information retrieval, queries are typically short, perhaps less than ten words.

Using either method, the greatest increase in performance for information retrieval comes from using a long query. Table 4 (below) shows some results produced by a probabilistic retrieval model. The document set consisted of AP news stories collected during April 1990. The query was the first April AP newswire story about Ryan White, a hemophiliac who died of AIDS on April 8, 1990. The table shows all stories in the collection which had a positive log likelihood score. That is, according to the model, these (and only these) stories are more likely to have been generated from the query story distribution than the global distribution.

In testing the stories, the headwords and titles were not used, so they can be examined as evidence of success in retrieval. It will be seen that the highest scoring stories are all relevant. The lower scoring stories tend to reflect part of the query, usually AIDS. No stories about Ryan White were omitted by this test. This example shows a remarkable performance for information retrieval, but it is due to using an entire story as the query, and not to the method. The vector space method does about as well on this example, although it does not have a natural cutoff, so comparison is difficult.

The information retrieval task shows that non-probabilistic methods may be better for some high dimensional problems, but that probabilistic methods can be applied "off the shelf" with competitive results.

Table 4: Probabilistic Search on "Obit-White" Story

loglike	date	headword	title
0.92	4-02	RyanWhite	AIDS Patient Ryan White Reported Near Death
0.36	4-02	RyanWhite	AIDS Victim Unconscious, Said to Be Dying from Internal ...
0.34	4-08	Bush-White	Bush Mourns Boy's Death
0.30	4-03	RyanWhite	
0.30	4-09	White-Chro	White's Struggle With AIDS Began At Age 13
0.21	4-03	RyanWhite	Ryan White, AIDS Patient, Near Death; Well-Wishers' Calls ...
0.20	4-03	RyanWhite	
0.18	4-04	RyanWhite	AIDS Patient Ryan White Remains in Critical Condition
0.18	4-09	Obit-White	Ryan White Taught Nation to Care About AIDS, Friends Say
0.18	4-04	RyanWhite	Town That Once Spurned Ryan White Joins Nation in Wishing
0.16	4-11	Bush-Health	Bush to Undergo Physical on Thursday
0.13	4-05	RyanWhite	In City That Barred AIDS Boy, an Outpouring of Sympathy
0.13	4-06	RyanWhite	Ryan White Brings Hope to Other Patients with Gifts from ...
0.13	4-10	RyanWhite	AIDS Victim Ryan White Remembered as Teacher, Student
0.12	4-07	Ryan'sLega	Unassuming Indiana Teen-ager Taught America About AIDS
0.11	4-10	RyanWhite	Hundreds Pay Respect at Funeral Home
0.11	4-11	RyanWhite	Barbara Bush, Celebrities to Attend Ryan White Funeral
0.09	4-11	RyanWhite	1,500 Say Goodbye to AIDS Victim Ryan White
0.09	4-08	White-Reax	Americans Pay Tribute to AIDS Victim Ryan White
0.09	4-12	RyanWhite	First Lady, Celebrities Attend Funeral for Young AIDS Victim
0.08	4-25	RyanWhite	Victim's Mother Lobbies for AIDS Bill
0.07	4-20	AIDSBoy	Church Bars AIDS-Infected Boy from Sunday School
0.06	4-20	DigestBriefs	
0.06	4-20	CDC-AIDS	Woman Gets AIDS Virus After Being Inseminated with Infected
0.06	4-20	AIDSBoy	Church Bars Boy With AIDS, Then Reverses Itself
0.06	4-10	Singapore	AIDS Test Required For New Maids
0.05	4-26	RyanWhite	AIDS Victim's Mother Lobbies for Spending Bill
0.05	4-22	Academic	Top 15 Finishers Listed
0.05	4-13	Students	Death of AIDS Victim Ryan White Sparks Protest
0.05	4-18	Scotus-Des	
0.04	4-04	RyanWhite	Hemophiliacs Live with Uncertainty of AIDS Infection
0.03	4-13	Kenya-AIDS	More than Four-Fifths of Kenyan Prostitutes Carry AIDS
0.03	4-16	RuralAIDS	AIDS Panel Studies Special Problems Of Rural Patients
0.03	4-05	Quotes	
0.03	4-09	Briefs	Lithuanian Declaration Should Be Withdrawn, Soviet Says

8 Capitalization Restoration

Although author identification and information retrieval are perhaps the two best known applications of high dimensional discrimination analysis in natural language processing, there are plenty of additional applications that are also worth mentioning. We have recently used these techniques for restoring capitalization in text that has been transmitted through a noisy channel that corrupted the case information. Some text, such as that collected from a teletype, does not preserve capitals. Also, in sentence initial position, all words are capitalized. Thus it is sometimes useful to be able to distinguish the two words by context. For many purposes, it is desirable to try to recover the capitalization information. For example, we may wish to be able to distinguish the country *Turkey* from the bird *turkey*.

We have made a few preliminary studies of this issue. It has been possible to gather large sets of training examples automatically, although considerable care was needed to avoid the multiplicity of situations in which all words are capitalized. The example of *[T/t]urkey* is a biased example: it is our most successful model to date. When the Bayesian model of context was trained on 200 examples each of *Turkey* and *turkey* drawn from the Associated Press newswire, and tested on an additional fifty examples of each, each of the 100 test examples was correctly classified. Performance is generally closer to 90% on more typical examples.

A similar problem, in which a large dimensional model would form part of the solution is the restoration of accents. The Spanish EFE newswire deletes accents, and hardware limitations in the immediate future may create other such situations. Accent deletion can create ambiguities. If some unaccented French text contained *peche*, for instance, one would need to distinguish among *pêche*, *pèche* and *péché*.

9 Person or Place?

Our final example is that of distinguishing proper nouns that refer to people from proper nouns that refer to places. *Madison*, for instance, can refer to the former president of the United States or a city in Wisconsin. Since our work on this question remains preliminary, we will keep the discussion brief.

For a preliminary study, we considered discriminating city names from people's names. City names are convenient because there are a number of readily available sources of city names (e.g., almanacs, atlases). We need to be careful in selecting training material. If we simply collected all examples that we happened to have lying around, we might well end up producing a model that would be heavily biased toward the frequent items and may not reflect the generality very well. Thus we trained using sets with one example each of each name.

The models were tested on sets containing a second example of each name. We found that for this problem, as opposed to the sense discrimination problem, a narrow context of just one or two words to either side was best. This provided us with a context model.

However, for this problem, there are often strong priors about whether a name represents a person or a city. *Bush* is never a place, and in contemporary text, it is unlikely that *Madison* will refer to the former president. For each of the names we had, we made an iterative calculation to reach a probability for its representing a person. The model classifies each instance of a name assuming person or city equally likely a priori; the fraction classified as people can be taken as a prior and the group reclassified. This is an Expectation-Maximization calculation (Dempster, 1977), known to converge to a local maximum, but of unknown relation to the global maximum, or the truth. We made a series of studies by Monte Carlo techniques, and derived a calibration curve.

About three fourths of the names had just person or just city examples in training material, and errors on these groups were less than one percent on test material. Table 5 shows an example of the performance for *Dixon*, a name that is quite ambiguous, with a nearly equal distribution between person and place.

The first column in the table is the computer's assignment. The double asterisk in this column shows the one of twenty five examples in which the computer assignment

Table 5: Person/Place Assignment

Guess	Score	Text
City	-495	The people of **Dixon** are reading about
City	-893	journalist will visit **Dixon** for a week
Person	1997	lighter sentence on **Dixon** because he had
Person	1998	attorneys for **Dixon** had filed a
City **	-647	trial would prevent **Dixon** from challenging
Person	49	Bailey denied **Dixon** his constitutional
Person	5449	Kunstler said **Dixon** contends the two
City	-2776	still standing in **Dixon** that Reagan att
Person	173	the savings bank **Dixon** owned , Vernon
City	-19	miles west to **Dixon** later Wednesday
City	-4594	Reagan moved to **Dixon** with his family
City	-1057	where proud **Dixon** residents brag ,
City	-4594	Reagan moved to **Dixon** with his family
City	-4347	lived in **Dixon** from 1920 until
City	-2063	years they called **Dixon** home .
City	-2711	family lived in **Dixon** and it's the
City	-3758	ntacted officials in **Dixon** in early 1987
City	-7736	he arrived in **Dixon** last Thursday ,
City	-2819	lived in several **Dixon** houses with his
Person	1315	official at whom **Dixon** and others in
Person	2788	estigation , which **Dixon** said could cont
Person	145	Jones credits **Dixon** with helping dev
Person	4944	Police said **Dixon** was shot in
Person	1901	a letter to **Dixon** asking if the
Person	690	sat motionless as **Dixon** read his ruling
		** = *Incorrect*

differed from that of one judge. The second column is the computer's score, negative for cities, positive for people. The final column is a a small amount of context so that the reader can judge the results. By combining priors with context evidence, we reach mid-nineties for accuracy in ambiguous cases, resulting in an overall error rate of less than two percent, since the ambiguous cases are only a quarter of the total.

This use shows the need for care in gathering training material, the selection of appropriate model parameters, the use of the model to construct priors, and the utility of combining priors with the contextual information. The study needs to be extended to other kinds of places besides cities, and it needs to deal explicitly with proper nouns that are new, probably by building models from internal clues (such as morphology (e.g., *-burg*)). Nevertheless, we are quite pleased with these preliminary results.

10 Can We Reduce the Dimensionality?

Indirect approaches to the problem of high dimensionality attempt to reduce the dimensionality.

Information retrieval work started in days when reducing dimensionality was essential to computing efficiency. Stop lists and term selection were developed as standard, if heuristic methods to reduce dimensionality. Now, however, large vocabularies can be dealt with mechanically, and the question asked of any dimension reduction scheme is how it effects precision and recall. One early method for dimension reduction was suffix removal, or stemming. Harman (1991) studied three different stemming algorithms in common use, including the simplest "S-stemmer" which just reduces English forms ending in s (plurals for nouns, third person singular for verbs) to their root, and two more complex schemes. She concluded "The use of the three general purpose stemming algorithms did not result in improvements in retrieval performance in the three test collections examined, as measured by classical evaluation techniques." Figure 1 may help explain why she came to this conclusion.

The quantities plotted in Figure 1 are the probabilities that a word is a noun given an apparently singular form (no s ending) or an apparently plural form (s ending). In the upper right corner, the figure shows a group of words, about 30 percent of this sample of fifty words, which are almost always nouns in either the singular or the plural. Even with these words included, the correlation is .67, low enough to show that there is considerably different usage between singular and plural forms. The remaining 70 percent of the words, however, have only a .37 correlation, showing very little relationship between usage of the singular and plural forms. Basically, the singular and plural forms usually represent dimensions which are not particularly parallel, so stemming is not much better than just dropping words arbitrarily in its dimension reduction. Conceivably, a careful study of particular nouns, locating those which were nearly always a noun in either the singular or plural forms could be used to reduce the dimensionality.

We made a brief study of the importance of words by frequency for sense tagging. We divided the Hansard vocabulary into ten groups based on frequency. The words in the first group were selected to be more frequent than words in the second, which were more frequent than those in the third, and so on. Each of the ten groups contained approximately the same number of *tokens*. Thus, the first group consisted of a small handful of high frequency words, whereas the tenth group consisted of a large number

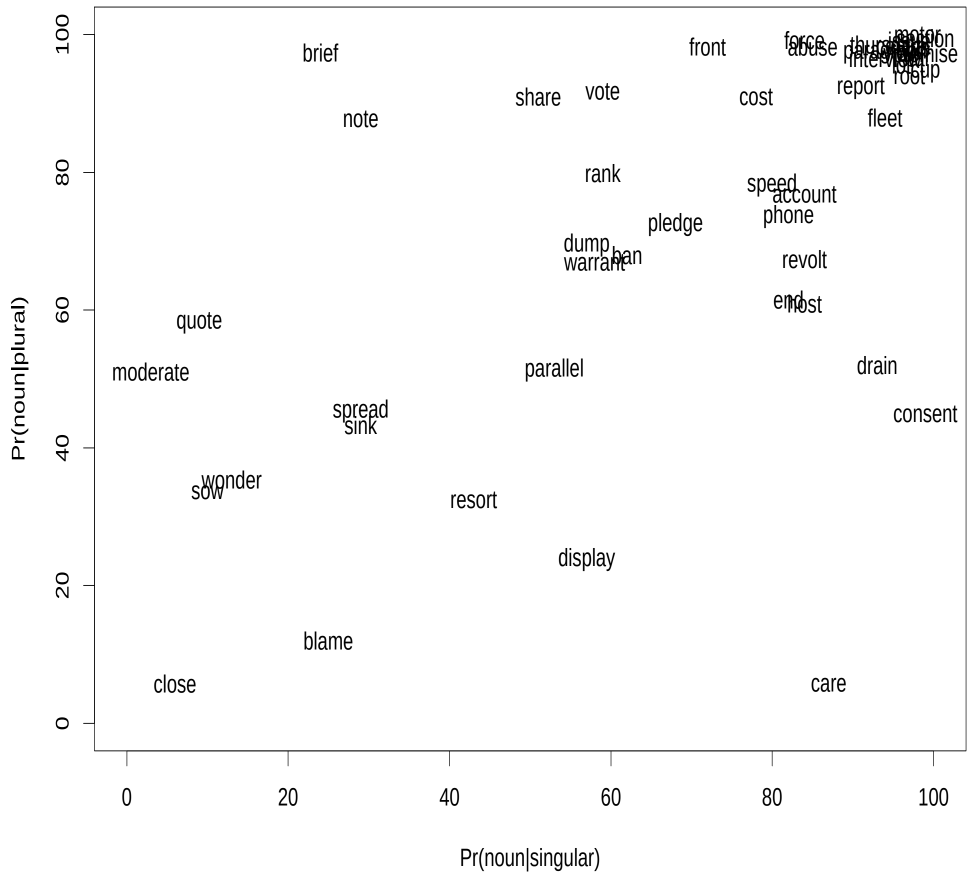

Figure 1: **Singulars and Plurals are Used Differently.** The horizontal axis shows the probability that a word will be a noun, rather than a verb or adjective, when seen in its base form (no added *s*). The vertical axis shows the probability that the same word will be a noun when it ends with the suffix *-s*. For example, in this study *brief* is rarely a noun (< 25%), while *briefs* is almost always a noun (> 95%). The correlation is .67, showing that the singular and plural forms are often used quite differently.

Table 6: Low Frequency Words
Discriminate Best

Group	Accuracy	Min. Freq.
1	.50	985,989
2	.57	587,847
3	.62	287,033
4	.67	134,531
5	.67	70,817
6	.73	24,309
7	.80	8,455
8	.84	3090
9	.88	802
10	.88	1

of low frequency words. We were interested in seeing how word performance would vary with word frequency.

Table 6 (above) shows the performance on the sense tagging application described earlier, averaged over the six polysemous words in Table 2. The third column shows the cut points defining the non-overlapping groups (a partition of the words) as the minimum frequency for the group. The frequencies are absolute frequencies out of a corpus with about 20 million words. Table 6 shows very clearly that the lower frequency groups out-perform the higher frequency groups.

Table 6 shows that the most desirable single groups are those with the most words. Thus restricting vocabulary by frequency will not reduce dimensionality. We also investigated the marginal utility of each frequency class, that is, the difference in accuracy due to adding one group to a model already based on all groups of lower frequency, beginning with a model based only on the group with the lowest frequency words. Each subsequent class, including surprisingly group 1, increased the accuracy of the models.

A group at Bell Communications Research has been investigating the use of singular value decomposition (SVD) as a means of dimension reduction. Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) discuss its application in information retrieval. As they say, "A fundamental deficiency of current information retrieval methods is that the words searchers use are often not the same as those by which the information they seek has been indexed." Singular value decomposition of the term by document matrix (the matrix with documents as rows, terms as columns, and the count of each term in a given document as the entry) selects directions in which groups of terms are found as dimensions, thus helping to solve this indexing-word problem, as well as reducing dimensionality. The authors report improvement in performance.

However, this is not yet an automatic solution to high dimension discrimination problems, because the size of the term by document matrix easily becomes unmanageable. The work cited used up to 2000 documents with up to 7000 terms. The previous discussion suggests that arbitrary reductions from 100,000 terms to 7000 terms will pay a price in accuracy. And while 2000 documents may be useful for some specialized collections, this is a small number when compared to the number of scientific papers published annually, say.

11 Significant Correlations are not Perfect Correlations

Our models, and most previous statistical models have assumed that the conditional correlations between terms are zero (the Bayesian coefficients are directly related to correlations between words and the selecting condition). In this section, "correlation" should be read as "conditional correlation." The most manageable alternative, used in stemming, is to assume that the correlation is perfect. The fundamental problem in considering interactions or correlations between the 100,000 dimensions of these models is that while we can manage $V = 10^5$ reasonably complex calculations, and have enough data to support the calculations, we have neither data nor disk space nor computer time to handle $V^2 = 10^{10}$ calculations. On the other hand, the interactions might be important for a number of our applications because there are many significant positive correlations among various pairs of words in the vocabulary.

Mosteller and Walace describe a theoretical model for adjusting a pair of words to account for their correlation (1968, section 4.7). They conclude "For our *Federalist* data the differences observed for the several methods suggest modest adjustments to the log odds." They make no adjustments to the log odds. Salton also discusses correlations, starting with his equation 10.18. He concludes "... not enough reliable term-occurrence information can be extracted from available document collections to produce improvements."

We have examined the question of dependency briefly for the sense discrimination problem. We follow the theory presented by Mosteller and Wallace. The results of their theory are as follows. Let the log odds due to word w_1 be γ_1^2, the log odds due to word w_2 alone be γ_2^2, and the correlation between the occurrences of w_1 and w_2 be ρ, and suppose without loss of generality that $\gamma_1 > \gamma_2$. Then if we knew just the evidence from w_1 we would have log odds of γ_1^2. The additional evidence from w_2 is

$$\frac{(\gamma_2 - \rho\gamma_1)^2}{1 - \rho^2}$$

Notice that for $\rho = 0$, the contribution is γ_2^2, as we currently assume. The derivation imposes a limit of γ_2/γ_1 for ρ. At this limit, the contribution from the second word is zero.

From the two models for duty, we selected the fifty most important words (maximum of score difference times frequency in the training sample), and calculated all pairwise correlations. A few were striking:

Table 7: Large Correlations		
0.36	tablets	tagamet
0.31	illicit	trafficking
0.28	organized	crime
0.24	prescription	patent
0.21	prescription	prices
0.21	tablets	valium

However, these were the only pairs exceeding .2 in correlation. Another 26 pairs exceeded .1 in correlation; none were below -0.1. The mean of the correlations was .01; the standard deviation of the mean was .001. In short, we agree with Mosteller and Wallace: the effects of correlations are modest.

12 Summary

We have discussed three major discrimination problems with large (100,000) dimensional spaces: sense discrimination, information retrieval, and author identification. The large number of dimensions results in each case from the number of terms each of whose frequencies will vary by context. We also gave two examples, capitalization restoration and person/place discrimination, from a much larger class of specific discrimination problems in high dimensional spaces.

Methods for these high dimensional spaces can basically be divided into two types: "direct," dealing with all the dimensions, or "indirect," attempting to reduce the number of dimensions first. We have shown that a Bayesian log odds model is a useful direct tool for each of the problems cited. It may not be the best tool for any of them, after thorough study, but it is easy to apply and gives results competitive with those of other methods where such other methods exist. It therefore appears to be a useful first cut tool for high dimensional discrimination problems.

There are problems with these methods. The Bayesian models have a problem of overstating their evidence because positive correlations between dimensions (words) cannot yet be accounted for. Brute force approaches to accounting for these correlations are not feasible, so some heuristics are needed. Also, as one example in author identification showed, the methods are currently not robust: evidence from one word can swamp the bulk of evidence from all other words. The methods need to be developed to overcome this undesirable feature, and used with caution until then.

While indirect methods have been necessary until recently due to computer limitations, the approaches have been heuristic. Since it has not been possible until recently to compare indirect methods with direct methods, little is actually known about the indirect methods. What little is known suggests they will be of limited utility or limited generality.

Acknowledgments

Discussions with Collin Mallows were vital to the development of the method presented here for the estimation of conditional probabilities. This paper has appeared in the *Annals of Operations Research.*

References

[1] Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer, "Word Sense Disambiguation Using Statistical Methods," *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 264-270, 1991.

[2] Church, K.W., "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Glasgow, 1989.

[3] Dagan, I., A. Ital and U. Schwall, "Two Languages are more Informative than One," *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 130-137, 1991.

[4] Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, 41, 1990.

[5] Dempster, A., N. Laird, and D. Rubin, "Maximum Likelhood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society (B)*, 39, 1977, pp. 1-38.

[6] DeRose, S. "Grammatical Category Disambiguation by Statistical Optimization," *Computational Linguistics*, 14, 1, 1988.

[7] Francis, W., and H. Kučera *Frequency Analysis of English Usage*, Houghton Mifflin Company, Boston, 1982.

[8] Gale, W., and K. Church, "Estimation Procedures for Language Context: Poor Estimates are Worse than None," pp. 69-74 in *Proceedings in Computational Statistics, 1990*, K. Momirivić and V. Mildner, eds., Physica-Verlag, Heidelberg, 1990.

[9] Gale, W., and K. Church "A Program for Aligning Sentences in Bilingual Corpora," *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991a, pp 177-184.

[10] Gale, W. and K. Church "Identifying Word Correspondences in Parallel Texts," *Proceedings of the DARPA Conference on Speech and Natural Language*, 1991b.

[11] Gale, W., K. Church, and D. Yarowsky "A Method for Disambiguating Word Senses in a Large Corpus," *Computers and the Humanities*, 26, 415-439, 1992.

[12] Harman, D., "How Effective is Suffixing?" *Journal of the American Society for Information Science*, 42, 1991, pp. 7-15.

[13] Harris, Z., *Mathematical Structures of Language*, Wiley, New York, 1968.

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

[15] Leacock, C., G. Miller, T. Towel and E. Voorhees, "Comparative Study of Statistical Methods for Sense Resolution," *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.

[16] Merialdo, B., 'Tagging Text with a Probabilistic Model," *Proceedings of the IBM Natural Language ITL*, Paris, France, 1990, pp. 161-172.

[17] Mosteller, Frederick, and David Wallace, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, MA, 1964.

[18] Salton, G., *Automatic Text Processing*, Addison-Wesley, Reading, MA, 1989.

[19] Salton, G. and C. Yang, "On the Specification of Term Values in Automatic Indexing," *Journal of Documentation*, 29, 1973, pp. 351-372.

[20] Yarowsky, D., "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora," *Proceedings of COLING-92*, Nantes, France, 1992.

[21] Yule, G. U., *Statistical Studies of Literary Vocabulary*, Cambridge University Press, Cambridge, England, 1944.

[22] Zernik, U., "Tagging Word Senses in a Corpus: The Needle in the Haystack Revisited," *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, P. Jacobs, ed., Lawrence Erlbaum, Hillsdale, NJ, 1992.

[23] Zipf, G. K., *Selected Studies of the Principle of Relative Frequency in Language*, Harvard University Press, Cambridge, MA, 1932.