Natural Language Processing In Support of Decision-Making: Phrases and Part-of-Speech Tagging

Information Processing & Management
37 (6), pp. 769-787, Nov. 2001

Robert M. Losee *
Manning Hall, CB#3360
U. of North Carolina-Chapel Hill
Chapel Hill, NC 27599-3360

losee@ils.unc.edu

July 11, 2001

Abstract

The use of natural language information can improve decision-making. Darwinian considerations suggest that language may have developed because it leads to improved decision making and survival, justifying the study of language's contribution to decision making. The study of information-based decision making within the context of evolution provides a view of information use that allows us to both describe the phenomenon of information use as well as to explain why an information use occurs as it does. Increasing information retrieval performance using phrases and part-of-speech (POS) information is one example of a type of decision-making performance that is improved when using this linguistic information. By studying a set of phrases used in a text retrieval system, we are able to show the relative effectiveness of using multi-term phrases as opposed to individual terms, as well as the relative worth of POS tagged terms or phrases, as opposed to untagged terms or phrases. An explanation is suggested for why POS tags contribute less to higher order grammatical constructs. We propose a measure of those needs for POS disambiguation that can be addressed by tagging; some example terms are analyzed using this measure, and specific degrees of ambiguity are proposed.

1 Introduction

By making decisions, and through rational (and irrational) processes, humans have developed complex social, political, and technical systems. Natural language was un-

*The author wishes to thank Miles Efron and an anonymous referee for their valued comments on an earlier draft of this article.

doubtedly contributed to the effective performance of these human-specific tasks. How, then, does the syntactic, semantic, or statistical information provided by natural language contribute to human decision-making?

Natural language has been studied using a wide range of techniques (Chomsky, 1965; Partee, Meulen, & Wall, 1990; Yngve, 1986; Groenink, 1997; Newmeyer, 1986; Oakes, 1998; Biber, Conrad, & Reppen, 1998). These may be roughly categorized, for our purposes, into *structural* linguistic methods and *statistical* methods.

Traditional methods assume that natural language may be described as a set of components that may be combined into progressively more complex components, with the components being understood to be based on conventional notions about language. These models assume that terms fall into specific categories, or parts-of-speech, and that terms are combined using grammatical rules to produce sentences (Chomsky, 1957, 1959).

The frequencies of term occurrences and their relationships to parts-of-speech of terms, as well as to the occurrences of other terms, may be modeled statistically. Chomsky, who argued in 1969 that "It must be recognized that the notion of a 'probability of a sentence' is an entirely useless one, under any interpretation of this term" (Ney, 1997), helped keep the assignment of part-of-speech tags away from the statistical. However, in recent years part-of-speech tagging has been moving back toward statistical methods as a result of recent successes (Biber et al., 1998; Manning & Schutze, 1999; Oakes, 1998).

But how to model language statistically remains an open question. Terms may be treated as individual entities that occur independent of the terms around them. Referred to as the *bag-of-terms* model, this approach simplifies models based on the frequencies of term occurrences. Term dependence models don't assume this separateness or independence of features; by incorporating term relationships, multivariate probabilities allow us to estimate the probability of occurrences for multi-term phrases. Term dependence models that have been used in pattern recognition and decision-making have included those that emphasize terms in trees (Chow & Liu, 1968; Van Rijsbergen, 1977), networks of terms including neural nets (Hopfield, 1982; Paass & Kindermann, 1995; Bishop, 1995) Bayesian inference networks (Heckerman, Geiger, & Chickering, 1995; Turtle & Croft, 1991), algebraic expansions incorporating dependences (Losee, 1994; Yu, Buckley, Lam, & Salton, 1983), and matrix methods (Losee, 1998; Teugels, 1990). Using these models of term relationships, we may improve decision-making at some cost to processing time and complexity.

2 Evolution and the Study of Natural Language

The ability to take beneficial actions that advantage a decision maker result in an evolutionary advantage for the decision maker (Altenberg, 1994). The ability for members of a species to communicate has evolved to produce a sophisticated body of language across a wide range of species, beautiful bird songs, and human language. Evolution changes a characteristic until it reaches a local "evolutionary equilibrium," the set of abilities or characteristics that can't be improved upon (Hirshleifer, 1982). Those not at an evolutionary equilibrium are less likely to survive than those at the (superior) evolutionary equilibrium. While some characteristics appear to be very localized (e.g. Kangaroos developed only in Australia) languages of various sorts have evolved in a number of ways and over a very wide range of species and may represent an evolutionary equilibrium for a number of species. While there is no rigorous proof that evolution has occurred or that particular features contributed to evolution, there are a wide range

of communicative phenomena that could have evolved and that can be argued to be beneficial. These are best summarized by Hauser (1996).

Natural languages have unique characteristics that can assist decision-making. Hockett has suggested that there are at least 13 different features (and thus functions) in human communication (Hauser, 1996). No other species has all 13 features, although many species have several. These unique characteristics support decision-making characteristics in humans that may be unique in the animal world.

Humans with specific linguistic capabilities can improve their decision-making and increase their chances for survival and reproduction. An individual may contribute to their own survival, or the survival of other humans by contributing to the increased chance of survival through improved actions. For example, someone who can communicate the nature of a vaccine for a childhood disease that can save the lives of millions of children may not increase their own chance of survival but will increase the number of communicative humans with intellectual abilities similar in some respects to those of the vaccine's developer (Losee, 1999). Terms are at or near the core of oral communication supporting decision-making.

Terms are often ambiguous. A simple examination of any dictionary provides evidence of the different semantic connotations for most terms. Through misunderstandings by the recipient of a communication, this ambiguity has an impact on human decision-making.

Removing ambiguity may be accomplished through a number of methods, described most fully in the recent special topic issue of *Computers and the Humanities* (Kilgarriff & Palmer, 2000). While using natural language processing techniques such as part-of-speech tagging may improve retrieval and filtering performance, the degree of improvement with word-sense disambiguation may vary from small to moderate amounts (Burgin & Dillon, 1992; Ide & Veronis, 1998; Krovetz & Croft, 1992; Sanderson, 1994; Strzalowski, 1995). Wilks and Stevenson (1998) found in a small test that "92% of content word tokens can be disambiguated .... using the part-of-speech information produced by a a part-of-speech tagger." Such tagging can be viewed as providing a rough approximation of what might be available through the application of more advanced artificial intelligence techniques (Kilgarriff & Palmer, 2000). Part-of-speech tagging disambiguates to some extent by separating terms that may have multiple meanings by taking advantage of information, for example, such as that fact that one group of term uses is identified as "nouns," while another group might be associated with the part-of-speech category "verbs." Clearly, being able to further separate terms into different meanings, for example, noun use 1, noun use 2, etc., will result in disambiguation beyond that sometimes offered when only using part-of-speech tags. These more nuanced forms of disambiguation may use semantic information, such as that provided when using a dictionary, or information may be provided by the context.

3 Decision-Making

Humans make decisions to improve their expected conditions in life. Benefits may be major consequences such as life-and-death, or they may be shorter term and less consequential, such as whether to eat an apple or a pear. Not all decisions appear to maximize long-term benefits: most humans occasionally make decisions that rational judgment would suggest are counterproductive but that have a strong sensory appeal, such as indulging in chocolate. Decisions involve taking those *actions* that are expected to maximize the benefit to the actor. These decisions are based on the costs or benefits to the user of particular actions.

Decision theory provides prescriptions, recommendations that should be followed by an economically rational individual. However, humans are not completely rational beings, and descriptive decision theory suggests how humans *do* make decisions. Below we will focus on how language benefits rational individuals making decisions; this is expected to characterize much of human decision-making, with prescriptive decision-making serving as an approximation of actual decision-making. It will clearly fail to capture the sub-rational processes that are a part of daily life (Kahneman, Slovic, & Tversky, 1982; Tversky, Slovic, & Sattah, 1988; Brenner, Koehler, Liberman, & Tversky, 1996; Russo, Medvec, & Meloy, 1996; Fishhoff, 1996).

A prescriptive model of decision-making assumes that the expected cost associated with an action should be minimized for the actor. This is consistent with the general decision rule:

> Choose an action if the expected cost of the action is less than the expected cost of not taking the action.

This also may be expressed as the following rule: take an action a only when

$$EC_a < EC_{\bar{a}},$$

that is, when the expected cost (EC) of performing action a is less than the expected cost of not performing action a, denoted as $\bar{a}$. Frequently a bar will be placed over a variable to indicate the negation of the variable or the variable being applied to the opposite action. Rules such as these are used widely in decision making (Pratt, Raiffa, & Schlaifer, 1995) and were introduced into the information retrieval field at an early date by Maron (1960).

The expected cost of performing an action is estimated probabilistically as

$$EC_a = \Pr(g|e)C_{a,g} + \Pr(\bar{g}|e)C_{a,\bar{g}},$$

where $\Pr(g|e)$ denotes the probability that the action results in a good outcome, given that it has evidence e, and the cost of performing the good action having result g is denoted as $C_{a,g}$, with $C_{a,\bar{g}}$ represents the cost associated with taking the action and producing the negative result $\bar{g}$. A similar expression is used for the expected cost of performing $\bar{a}$ with a bad consequence, $\bar{g}$.

We assume here two types of results: good (g) and bad ($\bar{g}$). This assumption is made for model simplification purposes. It may be the case that the quality of results is continuous, and that we can denote those results at or above a certain cutoff as *good* and those below the cutoff as *bad*. We can similarly use this "cutoff" model if we view the quality of actions as a small number of discrete values, with those at or above a cutoff value as *good* and those below as *bad*.

Note that each expected cost in our model represents the average of two costs. Each expected cost is the expected cost of an action, with the two costs being averaged between the two possible states-of-nature associated with the action. Thus the expected cost of taking the action a is the weighted average of the cost of taking the action with a good result g (astate-of-nature) and the cost of taking the action and having a bad result $\bar{g}$ (a state-of-nature).

We may modify our rule above to suggest that we should perform action a with evidence e if and only if

$$\Pr(g|e)C_{a,g} + \Pr(\bar{g}|e)C_{a,\bar{g}} < \Pr(g|e)C_{\bar{a},g} + \Pr(\bar{g}|e)C_{\bar{a},\bar{g}}.$$

This may be transformed to produce a rule suggesting that the action be taken if and

only if

$$\frac{\Pr(g|e)}{\Pr(\bar{g}|e)} > \frac{C_{a,\bar{g}} - C_{\bar{a},\bar{g}}}{C_{\bar{a},g} - C_{a,g}} = constant, \tag{1}$$

that is, action a should be taken if the odds that it is a good action, given the evidence e, exceeds a particular constant.

4 Information Retrieval as a Type of Decision-Making

Information retrieval is the discipline associated with the organization, ordering, and retrieval of documents to satisfy the information needs of end-users (Salton & McGill, 1983; Van Rijsbergen, 1979; Kowalski, 1998; Losee, 1998). Models of the decision to retrieve documents extend the basic decision theoretic model proposed above. When using Equation 1, one can decide whether to retrieve a document or not depending on whether the ratio on the left hand side of the equation exceeds the cost constant on the right hand side. In many retrieval situations, the user may find it difficult to provide the costs necessary to compute the cost constant, and the system *ranks* the documents by the value of the left hand side of the equation. The documents are then presented to the user in this ranked order. Ranking allows us to use this formula instead of determining the value for the costs.

Using the ratio on the left-hand-side of Equation 1 above, we find that

$$\frac{\Pr(g|e)}{\Pr(\bar{g}|e)} = \frac{\Pr(e|g)\ \Pr(g)}{\Pr(e|\bar{g})\ \Pr(\bar{g})}.$$

If we assume conditional term independence, that is, term frequencies in relevant and in non-relevant documents are statistically independent, we may then compute

$$\prod_{i=1}^{n} \frac{\Pr(E_i = e_i|g)}{\Pr(E_i = e_i|\bar{g})} \frac{\Pr(g)}{\Pr(\bar{g})}. \tag{2}$$

Assuming statistical independence of features is not a bad assumption if parameters are properly chosen (Wise, Perrin, Vaughan, & Yadrick, 1989).

We can drop the $\Pr(g)/\Pr(\bar{g})$ component in Equation 2, which is constant for any given query and independent of the characteristics of a particular document, to suggest that documents be ranked by

$$\prod_{i=1}^{n} \frac{\Pr(e_i|g)}{\Pr(e_i|\bar{g})}. \tag{3}$$

This formula can be used to rank documents with independent terms by first computing the probability of each term and thus each document having a particular feature value, given the appropriate relevance class, and then ranking the documents by this value.

We may estimate the probability of a term occurrence (evidence) e in relevant documents as

$$\Pr(e|rel, p) = p^e(1-p)^{(1-e)},$$

where p is the probability of a term occurring in a relevant document, and where the evidence e is assumed to be binary, with probability of occurrence of p. We will similarly treat the same distribution for non-relevant documents, where a term is present

with probability q, the probability that a non-relevant document has a term. Equation 3 becomes

$$\prod_{i=1}^{n} \frac{\Pr(e_i|g)}{\Pr(e_i|\bar{g})} = \prod_{i=1}^{n} \left(\frac{p_i/(1-p_i)}{q_i/(1-q_i)} \right)^{e_i}.$$

Based on this model, queries may be modified either through adding new terms, query expansion (Efthimiadis, 1996; Lu & Keefer, 1995), or by treating the original query as a vector of all terms, with those terms in the query presented to the system as having relatively higher weights than those terms not included in the presented query (Losee, 1988, 1998). Relevance feedback can then be used to modify the term weights based upon user statements about which documents the user found helpful and those that weren't judged helpful.

If we limit ourselves to the analysis of a single term or a grammatical multi-term phrase, we can predict the expected retrieval performance. For our purposes in a term matching system, such as those described above, only a single feature in the query is considered. While this single term model may appear very simple, it will allow for sophisticated analysis of term relationships and part-of-speech tagging.

We denote the probability that the single feature occurs in relevant documents as p or unconditionally, in any document at all, as t. Given an optimal ranking method, the Average Search Length (ASL), the expected position of a relevant document in the list of ranked documents, is N(1-p+t)/2 + 1/2. For a given set of data, one may compare different aspects of retrieval, such as part-of-speech tagging, by comparing $\mathcal{A} = (1 - p + t)/2$ values. Here $\mathcal{A}$ represents one minus the proportion of documents occurring before the average position of a relevant document in an optimally ranked list of documents (Losee, 1998). One may compute the average search length as $ASL = N\mathcal{A} + 1/2$ when ranking is optimal and there are N documents. The best case ASL performance approaches 1 and the worse case approaches N documents. $\mathcal{A}$ is a function of the data being ranked and the features used, and not of the ranking algorithm (Losee & Paris, 1999). Below, we will use $\mathcal{A}$ to compare the ranking that is possible, assuming the existence of an optimal ranking method.

5 Experimental Data and Measurement

We study how phrases and their associated term dependencies, as well as parts-of-speech, contribute to decision-making by examining retrieval performance in a standard dataset (Fagan, 1989; Losee, 1994, 1996b). The *Cystic Fibrosis* (CF) database was developed by locating all documents in the U.S. Government's National Library of Medicine Medlars database that contained the subject heading *Cystic Fibrosis* and that were entered between 1974 and 1979 (Shaw, Wood, Wood, & Tibbo, 1991). Each document was judged for relevance for each of 100 queries by a team of medical specialists in the area of Cystic Fibrosis. This database, with its exhaustive relevance judgments, is considered a high quality retrieval database. A subset of this database has been developed containing the fulltext of about one third of the medical articles; this subset was used for our study (Moon, 1993).

Terms and phrases are included in our study when they occur in both the document and the query. In cases where parts-of-speech are being considered, the parts-of-speech for each term in the document phrase must match the parts-of-speech found in the terms in a matching query phrase. These tags may be assumed to evolve over time; however, here they are treated as a predetermined set (Lankhorst, 1995; Losee, 1996a). Part-of-speech tagging of terms in queries and documents was done using the Brill part-of-

speech tagger (Brill, 1994). The tags described below are those used by Brill and the Penn Treebank (http://www.cis.upenn.edu/~treebank/), or minor variants of them. Brill claims a tagging accuracy of over 96% for his test corpus, with a surprisingly high accuracy of 85% for unknown terms. Since the tagging used in this study was based on Brill's training, the accuracy of the tagging for the CF database would probably be somewhere within this range of values.

NOTE: Tilde before "treebank"

Because our model addresses the binary presence of phrases and terms, we note whether a phrase and part-of-speech combination simply occurs in a document; the frequency of occurrence is ignored for the data analysis here. $\mathcal{A}$ values were computed for the situation where each query contained the phrase in question and the documents contained (or lacked) the phrase being considered. We are thus able to study the efficacy of each type of phrase that supports decision-making.

Values are given for $\mathcal{A}$ for two different linguistic situations. The values labeled $\mathcal{TP}$ are performance averages with term and part-of-speech information used in the retrieval. The values labeled $\mathcal{T}$ are averages computed with only the term (and not with the part-of-speech tags) being used in retrieval.

Averages given below were computed by taking $\mathcal{A}$ values for each query (for the appropriate data) and then averaging the $\mathcal{A}$ values over all the queries. In the case of part-of-speech variables, the averages are for the parts-of-speech; the averages are thus for the part-of-speech types, not the tokens, the individual term occurrences. This allows us to gain a better understanding of the variance over the parts-of-speech (Biber, 1988). Our goal is to provide recommendations about when part-of-speech information should be used, and to a lesser extent to provide evidence about the number of different types (*not* tokens) that result in improved performance.

One factor that may affect decision-making performance and the results presented below are the nature of expressed queries. The 100 queries that are part of the standard Cystic Fibrosis database are what individuals who produced the queries considered to be typical queries. Queries presented to librarians and to information retrieval systems often tend to be brief and topical in nature. An example of this can be seen by looking at one of the search engine "spies" available on the World Wide Web (e.g., http://www.metaspy.com). Watching these for a few minutes shows that searchers often enter nouns or noun phrases. These searches often tend to be one or two terms in length. Further study will need to be done as decision-making systems accept greater amounts of information as queries. The tendency to give brief, telegraphic, queries may change when providing large quantities of data to a decision support system becomes relatively easy. Imagine being able to explain an entire problem to a system using voice input. In a matter of 60 seconds, the system would have a far richer query than the few terms used now, containing an order of magnitude more terms and depth than are found in most queries presented to systems today.

6 Decision Support Provided by Individual Terms

There are several different types of $\mathcal{A}$ performance relationships that can exist between single terms, both with and without their associated parts of speech attached as labels. We denote the $\mathcal{A}$ performance of terms or phrases with $\mathcal{T}$ and the terms or phrases combined with their parts of speech are denoted as $\mathcal{TP}$. A set of these values for the CF database is given in Table 1. Note that in the table, N indicates the number of times the phrase type occurs in a query and in either any of the relevant documents or in any of the non-relevant documents. Given the two relevance classes of documents, the number of phrases occurring in the 100 queries is approximately $N/2$.

Part-of-Speech	$\mathcal{TP}$	$\mathcal{T}$	N
CC	0.565	0.588	3
DT	0.652	0.579	9
IN	0.605	0.555	20
JJ	0.406	0.428	120
NN	0.360	0.411	175
NNP	0.448	0.417	223
NNPS	0.500	0.453	9
NNS	0.409	0.430	93
RB	0.506	0.515	31
VB	0.447	0.410	52
VBD	0.522	0.487	17
VBG	0.441	0.407	26
VBN	0.512	0.526	19
VBP	0.466	0.438	37
VBZ	0.498	0.451	17
WP	0.579	0.599	3

Table 1: $\mathcal{A}$ performance values for specific parts-of-speech.

Examining part-of-speech categories in Table 1 shows that some part-of-speech categories, such as common nouns (NN), are generally good discriminators and increase decision-making performance. Members of other part-of-speech categories, such as determiners (DT), are weak or negative discriminators. While this may appear to be a straw-man argument, this does provide concrete evidence that, as one would expect, these terms are poor discriminators. Using these negative discriminators can result in performance somewhat worse than random. Terms and their parts-of-speech clearly have a strong impact on retrieval performance and decision-making.

Adjectives (JJ) and common nouns (NN) are better discriminators than average. For both types of terms, the combination of term and part-of-speech together result in the better performance (lower $\mathcal{A}$) with part-of-speech information attached than without. We denote two $\mathcal{A}$ values x and y that are close or similar with $x \simeq y$, and a value x that is somewhat lower (better) than another, y, with $x \preceq y$. Thus, we may denote the relationships in adjectives and common nouns as $\mathcal{TP} \preceq \mathcal{T}$. The performance may be better for these two types of terms if the part-of-speech tag is attached because of the increased disambiguation that results due to the part-of-speech tagging.

Probably the simplest performance relationship occurs when $\mathcal{TP} \simeq \mathcal{T}$. In this case, the part-of-speech information adds little to the decision-making performance when using the term. An example of this in Table 1 is RB, an adverb. Such a term has a performance value close to random, and the part-of-speech labeling of such a phrase adds little information that would support decision-making.

In the case where $\mathcal{T} \preceq \mathcal{TP}$ the term or phrase performance is quantitatively better (has a lower $\mathcal{A}$ value) than that provided by the term with the part-of-speech information. A number of the verbal forms in Table 1, as well as singular proper nouns, NNP, show that the best performance is obtained with the untagged term.

In the case of proper nouns, the failure of tagging to improve performance may be because of typographically inconsistent data or erroneous tagging by the Brill part-of-speech tagger. A term in a query that has the first character in upper-case and a document with all characters for the same phrase in all lower case might be identified as different terms if they have part-of-speech tags attached. In the CF database, for

	Parts-of-Speech	$\mathcal{TP}$	$\mathcal{T}$	N
Descriptive	JJ_NN	.339	.382	40
data	NN_JJ	.384	.458	4
	NN	.360	.411	175
	JJ	.406	.428	120
Differences	NN -(JJ_NN)	.021	.029	
based on POS	JJ - (JJ_NN)	.067	.036	
Decrease in $\mathcal{A}$	(JJ_NN) - $\mathcal{TP}_{JJ\_NN}$	—	.043	
from upper-bounds	NN - $\mathcal{TP}_{JJ\_NN}$	.021	.072	
$\mathcal{TP}_{JJ\_NN}$	JJ - $\mathcal{TP}_{JJ\_NN}$	.067	.089	
Differences	NN -(NN_JJ)	-.024	-.047	
based on POS	JJ - (NN_JJ)	.022	-.030	
Decrease in $\mathcal{A}$	(NN_JJ) - $\mathcal{TP}_{NN\_JJ}$	—	.074	
from upper-bounds	NN - $\mathcal{TP}_{NN\_JJ}$	-.024	.027	
$\mathcal{TP}_{NN\_JJ}$	JJ - $\mathcal{TP}_{NN\_JJ}$	0.22	.044	

Table 2: Performance values with adjective noun (JJ_NN) and noun adjective (NN_JJ) phrases and performance of their components.

example, *Cystic Fibrosis* is treated as two proper nouns by the Brill tagger, while *cystic fibrosis* is treated as an adjective followed by a simple noun. Clearly, tagging here could lead to poor matching behavior.

It is less obvious why several of the verbal types (e.g. VB, VBG, VBP, and VBZ) show that performance is better without tagging. This might be due to parser errors, but may also be due to the presence of such verbal expressions in queries being somewhat misleading in terms of matching queries with relevant documents.

Terms from many part-of-speech categories, such as simple nouns, appear to be positive discriminators and aid in decision-making. Other parts-of-speech, such as determiners, contribute noise to the matching process.

7 The Construction of Phrases

While single terms can be used for matching queries and documents, the use of phrases is commonly expected to result in better decision-making performance because phrases often represent more complex and nuanced concepts and relationships than do single terms, whether the single terms are taken individually or whether they are brought together in an unordered grouping. When we discuss the performance of a specific type of phrase or, for example, all phrases of length 4, we will compute performance from the phrase as a whole, not as the additive performance that would be obtained by combining the measured performance levels of each individual component in a phrase.

One positively discriminating phrase is the adjective noun complex (JJ_NN). Given the discrimination power of simple nouns (NN) by themselves (Table 2 and Figure 1), one would expect that JJ_NN phrases would be good discriminators. Performance is better for the phrase as a whole than for either of the term types taken separately, whether we apply tags to the whole and the parts or whether we don't apply tags to the whole and the parts.

The fifth and sixth lines of Table 2 show the difference in performance, given a specific method, between the phrase as a whole and the individual terms. The seventh

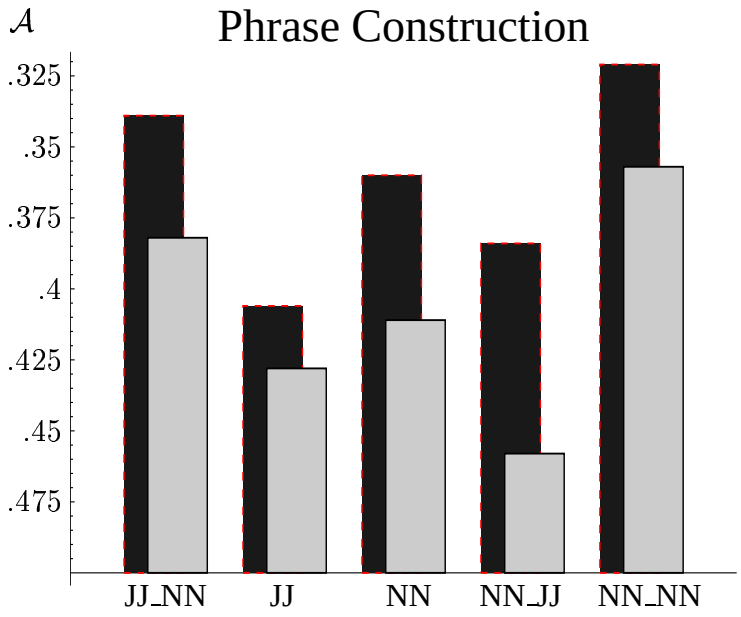

Figure 1: Performance for an adjective noun phrase (JJ_NN), noun adjective phrase (NN_JJ), and noun noun phrase (NN_NN). Lighter bars represent performance without part-of-speech tags ($\mathcal{T}$) and darker color bars show performance with the part-of-speech tags ($\mathcal{TP}$.) Lower values for $\mathcal{A}$ at the top of the graph represent better performance than higher values for $\mathcal{A}$ at the bottom of the graph.

	Parts-of-Speech	$\mathcal{TP}$	$\mathcal{T}$	N
Descriptive	NN_NN	.321	.357	24
data	NN	.360	.411	175
POS difference	NN - (NN_NN)	.041	.054	
Decrease in $\mathcal{A}$	NN_NN - $\mathcal{TP}_{NN\_NN}$	-	.036	
from upper-bounds	NN - $\mathcal{TP}_{NN\_NN}$	.041	.090	

Table 3: Performance values for noun_noun (NN-NN) phrases and their components.

through ninth lines show how each term or phrase, tagged or not, compares with the upper-bounds, $\mathcal{TP}_{JJ\_NN}$. These performance values are ordered in terms of decreasing performance (increasing $\mathcal{A}$)

$$\mathcal{TP}_{JJ\_NN} \preceq \mathcal{TP}_{NN} \preceq \mathcal{T}_{JJ\_NN} \preceq \mathcal{TP}_{JJ} \preceq \mathcal{T}_{NN} \preceq \mathcal{T}_{JJ}.$$

Clearly tagging is superior to not tagging for relatively highly discriminating phrases such as JJ_NN. In addition, it is clear that the phrase as a whole results in better decision-making than do the individual components.

Looking at the bottom two sections of Table 2 shows that the noun adjective (NN_JJ) phrase performs worse than do nouns alone. This phenomenon, where the phrase as a whole doesn't discriminate as well, on the average, as do the components of the phrase when taken alone, is indicative of a an anomalous structure which should be considered for exclusion from use in decision-making.

Another phrase with positive discriminating characteristics is shown in Table 3. Two common nouns occurring in sequence (NN_NN) are better discriminators, taken as a unit, than is a single noun by itself. The different term and tagging options show the following relationships between decision-making performance:

$$\mathcal{TP}_{NN\_NN} \preceq \mathcal{T}_{NN\_NN} \preceq \mathcal{TP}_{NN} \preceq \mathcal{T}_{NN}.$$

As with the JJ_NN phrase, the NN_NN discriminates better when tagged than do its components, whether tagged or not.

Fagan concluded that identifying just those phrases likely to be important appears to be difficult using quantitative methods alone (Fagan, 1989). We believe that methods based on part-of-speech tagging can begin to determine those phrases likely to be important, as well as those likely to be neutral or negative.

8 Phrase Length and Part-of-Speech Tags Improving Decision-Making

Part-of-speech tags clearly have the potential to improve retrieval performance, as was shown in the previous section. Here we examine performance averages using part-of-speech tags and attempt to make some generalizations about part-of-speech tags and the number of terms in a phrase.

Using decision rules developed in Losee (1998), we may analyze the phrases and associated tags from the CF database. We begin our examination by specifying the conditions under which part-of-speech tagging improves performance. We denote the probability that the query term occurs in a document as t, and the probability that a

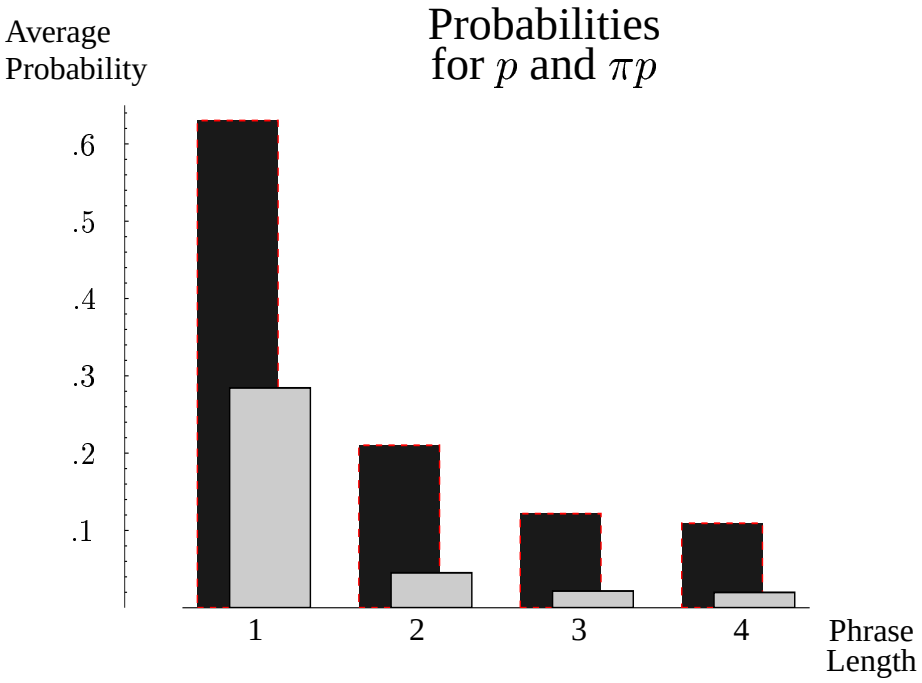

Figure 2: Values p (dark) and πp (lighter) for different phrase lengths.

document has the query term, given that the document is relevant, as p. The probability that a document is tagged with the same part-of-speech tag as that assigned to the term in the query, given that the query has the term, is denoted as τ. The probability that a document has the query term and is tagged with the query tag is the product $t\tau$. Similarly, the probability that a term is tagged with the query tag, given that the document has the term and is relevant, is π. The probability that a relevant document contains the term tagged with the query term's tag is the product $p\pi$.

Given these variables and optimal ranking, performance is improved if and only if

$$1 + t\tau - p\pi < 1 + t - p. \tag{4}$$

Here the left hand side represents $\mathcal{A}$ with part-of-speech tagging and the right hand side $\mathcal{A}$ without part-of-speech tagging.

Figures 2 – 4 show how tagging affects decision-making for decreasing phrase lengths, based on the relationships between term probabilities and tagging for relevant and for all documents. Figure 2 shows the relationship between the average p and πp. Similarly, Figure 3 shows the relationship between the average t and τt values.

How often does part-of-speech Tagging Help?

Individual terms or all those terms of a particular part-of-speech may be examined to see how often part-of-speech tagging improves decision-making. Using all terms of a particular part-of-speech and applying Equation 4, we can determine the percent of parts-of-speech that have improved performance when tags are applied. As is shown in Figure 4, the probability that a part-of-speech tag improves performance decreases

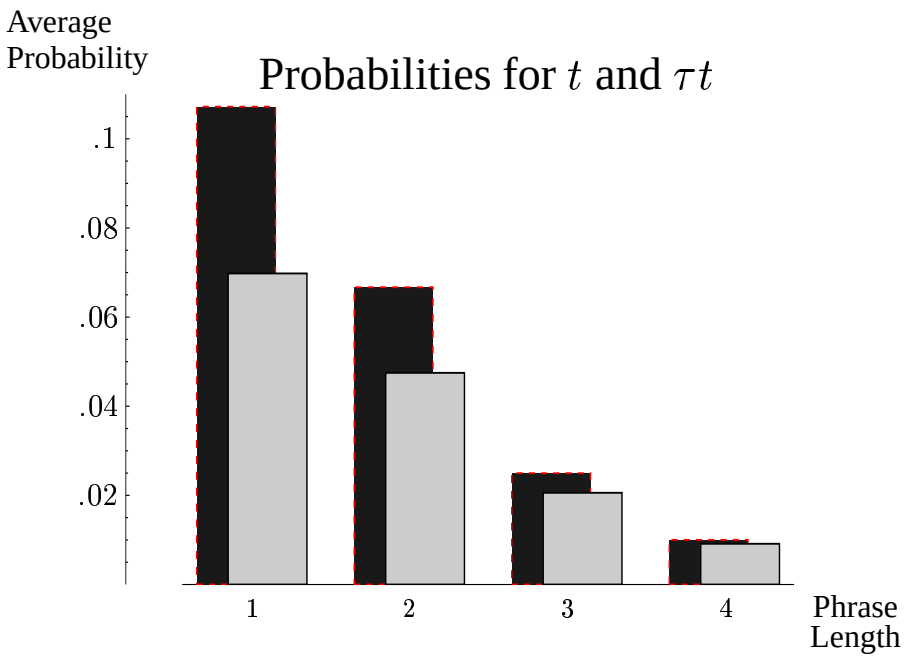

Figure 3: Values t (dark) and τt (lighter) as phrase length varies.

in an almost linear manner as the phrase length increases. As we move toward larger phrases, we move toward more complex structures which are intrinsically weaker at carrying the "aboutness" in the query than are what might be considered "stronger," single terms. A single tagged term type has an almost even chance of being a positive discriminator. These terms are useful for simple matching. As phrases grow in length, we find that the average phrase structures carry fewer nuances that match exactly with the concepts' phrase types in the query.

As one examines longer query phrases in Figure 2, one finds that πp decreases rapidly at first, and then decreases at a much slower rate. Interestingly the gap between p and πp decreases at a slower and steadier rate. Figure 3 shows the gap between t and τt steadily decreasing, to the point where it is almost nonexistent. The steady decrease seen in Figure 4 in the probability that a part-of-speech tag will improve performance as queries increase in length appears to be a combination of the relatively steady decrease in τt and the initially-rapid decrease in πp.

The downward trend in the probability a phrase type discriminates as the number of terms in a phrase increases shows that part-of-speech tags add less and less as phrases grow in length. Yet, we know that phrases can be superior to single terms at carrying content. The additional discrimination value in phrases must be due to structural and semantic information and not to part-of-speech tagging.

The performance with single terms is much better for relevant documents than for all documents because of the discrimination power of several specific term types, as was shown in Table 1. As one increases phrase size, from the smaller to the larger structures found in queries in Figures 2 to 4, we find that, on the average, larger, multi-term phrase types don't discriminate as well as do the smaller, single term types.

One partial explanation for the relationship between phrase length and the degree

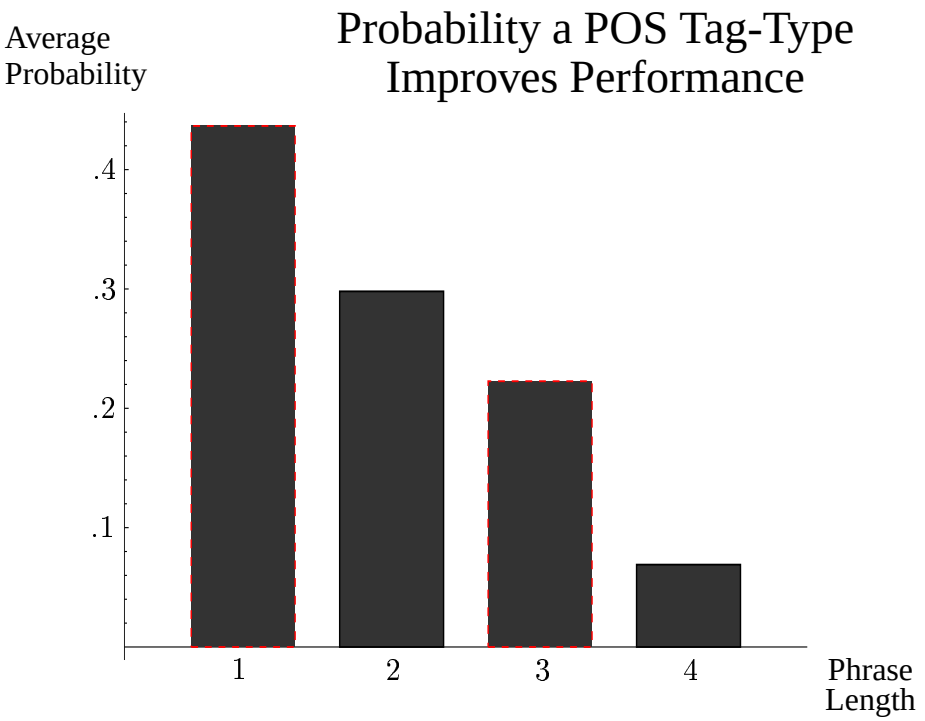

Figure 4: Increasing phrase length correlates with decreasing probability that a part-of-speech tag type improves retrieval performance.

to which part-of-speech tagging improves discrimination performance may be based on the evolution of language and linguistic capabilities in humans. If natural language grammars and parts-of-speech developed initially with the shorter speech segments that undoubtedly existed before long speech segments, long before people became verbose and began stringing large numbers of terms together, one would expect part-of-speech information to be effective for short linguistic segments. However, as language became more precise and sophisticated, one can imagine the language not requiring part-of-speech tags to the extent that they are required with shorter speech segments. Thus, tagging would yield better marginal discrimination with shorter phrases than with longer phrases. In addition, as meaning becomes expressed in increasingly long phrases, the individual terms lose more and more of their identity and power to discriminate and this power is taken on by the larger phrases.

9 Upper-Bounds for Performance with Tagging

Can the part-of-speech tagging produce better disambiguating performance than has been reported here? Retrieval performance may be viewed as being dominated by t and p (Losee, 1998). Part-of-speech tagging may be viewed as providing a means of improving performance by modifying these values through τ and π. However, part-of-speech tagging cannot overcome some performance constraints imposed by the untagged probabilities p and t.

The upper-bounds for performance without tagging may be estimated from the situation where p is at its maximum and t at its minimum, approaching 0, producing a

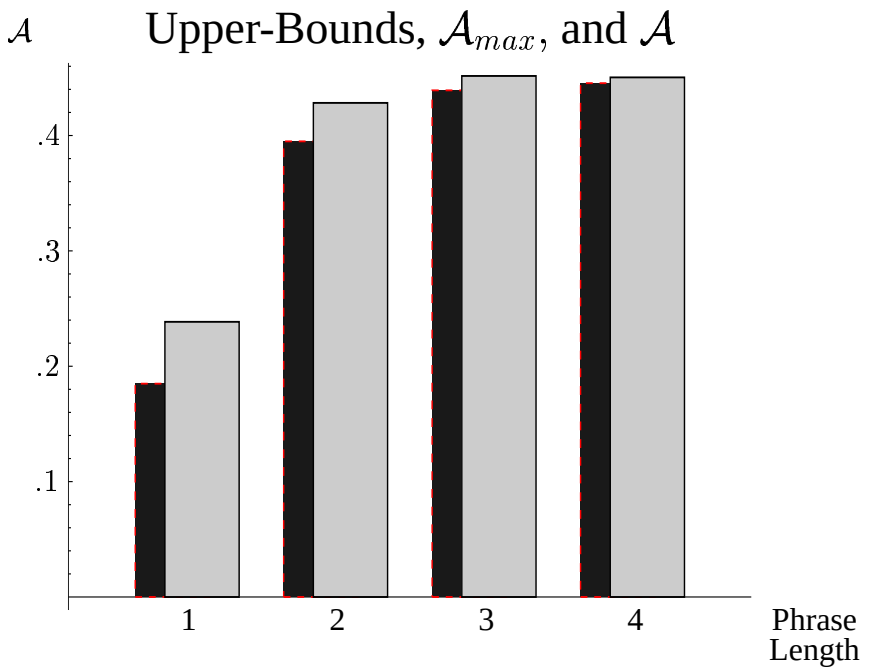

Figure 5: Average upper-bounds, $\mathcal{A}_{max}$ (dark), and $\mathcal{A}$ (lighter) for phrases of varying lengths.

maximum $\mathcal{A}$ value that approaches $(1-p)/2$. The exact upper performance bounds with part-of-speech tagging and generality $g = \Pr(rel)$ are

$$\mathcal{A}_{max} = (1 + gp - p)/2 \tag{5}$$

This assumes that $\pi = 1$ and τ approaches 1. We assume that there are at least (gp) tagged terms in the document database. The gp values in Equation 5 become very small, approaching 0 when a very large, realistic database is used, so that this equation in the limiting case approaches $(1-p)/2$. We use this estimate in computations below.

In most realistic searches, t will be rather small if it is a good search term, usually below .01. When p is much higher than this, as would be the case with a strongly discriminating term, the potential for improvement with tagging is far better than the potential for a decrease in performance.

As phrases from queries lengthen, both $\mathcal{A}$ and the upper bounds, $\mathcal{A}_{max}$, increase (and performance decreases), as is shown in Figure 5. The gap between $\mathcal{A}$ and the upper-bounds $\mathcal{A}_{max}$ decreases as phrases lengthen.

A comparison of Figures 2 and 3 shows πp decreasing rapidly, but not as rapidly as τt decreases. This accounts for the decrease in performance (and increase in $\mathcal{A}$) as we move toward longer queries. The sharp change in $\mathcal{A}$ moving from a query phrase length of 1 to a length of 2 is due largely to the sharp drop in πp moving from query phrase length 1 to 2.

Similarly, the upper-bound $\mathcal{A}_{max}$ increases (and performance decreases) at a faster rate than does $\mathcal{A}$. This can be explained by examining the p values, which drop more sharply as query phrases lengthen than do the πp values.

10 Measuring Part-of-Speech Disambiguation Capabilities

Disambiguating terms and phrases may take two forms. We refer to the disambiguation due to part-of-speech tagging as *part-of-speech disambiguation.* When the term *run* is labeled as a verb, with several meanings, or as a noun, with many meanings, the term has been *part-of-speech disambiguated*. *Non-part-of-speech disambiguation*, on the other hand, represents the word-sense disambiguation that fully separates different uses of terms but is not provided by part-of-speech tagging. The term *run*, for example, has several different meanings as a noun, and part-of-speech tagging doesn't help us disambiguate these meanings. Clearly, performing non-part-of-speech-disambiguation requires either additional semantic or dictionary knowledge (from a traditional linguistics perspective) or additional dependence information available from the context (from the corpus or statistical perspective) (Biber et al., 1998; Ide & Veronis, 1998; Kilgarriff & Palmer, 2000; Losee, 1994; Oakes, 1998).

Some measures of the disambiguation of specific methods are based on the number or percent of terms that are ambiguous, or unambiguous, given different conditions (Sutcliffe & Slater, 1995). We may measure the part-of-speech disambiguation capability associated with a term as $\mathcal{D}_{all}$, one minus the weighted average of the $\Pr(tag|term)$ values taken over the set of possible tags,

$$\mathcal{D}_{all} = 1 - \sum_i \Pr(tag_i|term)^2. \tag{6}$$

The sum is taken over all possible tags. When a term occurs with only a single part-of-speech, the probability in the expression is 1 and the part-of-speech disambiguating capability $\mathcal{D}_{all}$ is 0, while if the term is spread over a very large number of parts-of-speech, the probability will be very small and the $\mathcal{D}_{all}$ approaches 1.

Clearly, situations where $\mathcal{D}_{all}$ is high are those cases where part-of-speech disambiguation adds the most to decision-making performance. In an operational system that has the option of tagging or not tagging, one criteria for choosing to tag terms might be the average value of $\mathcal{D}_{all}$.

An analysis of terms and phrases that occurred in both queries and documents in the CF database shows $\mathcal{D}_{all}$ values ranging from 0 to about 0.5. We use as the cutoffs and labels for these part-of-speech levels:

- $\mathcal{D}_{all} = 0$, *unambiguous*;
- $0.03 \geq \mathcal{D}_{all} > 0$, *minimally ambiguous*;
- $0.3 \geq \mathcal{D}_{all} > 0.03$, *moderately ambiguous*; and
- $\mathcal{D}_{all} > 0.3$, *highly ambiguous*.

Unambiguous terms such as *mucus* and *metabolism* had $\mathcal{D}_{all} = 0$. *Minimally ambiguous* terms, such as *cf* and *role* had $\mathcal{D}_{all}$ values below 0.03 but above 0. Examples of *moderately ambiguous* terms include *patient* ($\mathcal{D}_{all} = .153$) and *test* ($\mathcal{D}_{all} = .190$). Terms considered *highly ambiguous* are those such as *influence* ($\mathcal{D}_{all} = .572$) which has tags of NN, NNP, VBP, NNS, and VB. This measure of ambiguity allows us to formally discuss ambiguity and the abilty to part-of-speech disambiguate, as well as to compare the relative ambiguity of different terms.

11 Discussion

Decision-making is improved by using part-of-speech tags and phrases in many cases. Positive decision-making performance is obtained by treating queries presented to retrieval systems as a bag-of-terms, a set of statistically independent terms. The data presented above and in the term dependence literature suggest that terms brought together and treated as phrases will outperform independent term models and that phrases often, but not always, are better than single terms taken alone. Similarly, tagging terms with part-of-speech tags will improve decision-making performance in many, but not all, cases. Increasing the length of the phrase type that is being matched in the query and the document results in less of a performance gain than with shorter phrase types. This occurs because of the greater complexity of longer phrases, although there is still a performance gain. It appears that much of the performance gain in simple term matching is due to individual terms, although there is a performance gain with multiple terms.

Based on empirical tests, we can address how the performance obtained with larger groupings of untagged terms compares with those same groups when they are tagged. As we see in Figure 5, the $\mathcal{A}$ values based on the average probabilities decrease in performance as query phrases lengthen. The change in probabilities of occurrences for p, πp, t, and τt in Figures 2 and 3 produce the decrease found in Figure 4. The difference between Figure 2 and Figure 3 suggest that the part-of-speech tagging functions somewhat differently in relevant documents than in all documents. The data in Table 1 shows that, for many parts-of-speech, tagging improves performance of a decision-making system.

Given these results, there are several recommendations that can be made for those developing decision-making systems. Part-of-speech tagging should be performed if not too expensive and no other disambiguating forms of information are available that would lead to semantic differentiation between ambiguous terms. In addition, methods based on part-of-speech tagging were proposed to determine those phrases likely to be important, as well as those likely to be noisy. This can be used to isolate useful phrases and to exclude neutral or negative discriminating types of phrases. For example, the results above suggest that adjective noun phrases are, on the average, stronger discriminators than most other phrase types. Retrieval systems that can isolate these discriminating phrases might use them in addition to individual terms in a bag-of-terms model.

References

Altenberg, L. (1994). The evolution of evolvability in genetic programming. In Kinnear, K. E. (Ed.), *Advances in Genetic Programming*, pp. 47–74. MIT Press, Cambridge, Mass.

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge Univ. Press, Cambridge.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge Univ. Press, Cambridge, U.K.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press, New York.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes, 65*(3), 212–219.

Brill, E. (1994). Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, pp. 722–727 Menlo Park, CA. AAAI Press.

Burgin, R., & Dillon, M. (1992). Improving disambiguation in FASIT. *Journal of the American Society for Information Science, 43*, 101–114.

Chomsky, N. (1957). *Syntactic Structures.* Mouton, The Hague.

Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control, 2*(2), 137–167.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT Press, Cambridge, Mass.

Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory, IT-14*(3), 462–467.

Efthimiadis, E. N. (1996). Query expansion. In *Annual Review of Information Science and Technology,* pp. 121–187. Information Today, Inc., Medford, NJ.

Fagan, J. L. (1989). The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science, 40*(2), 115–132.

Fishhoff, B. (1996). The real world: What good is it. *Organizational Behavior and Human Decision Processes, 65*(3), 232–248.

Groenink, A. (1997). *Surface Without Structure.* Ph.D. thesis, University of Utrecht, Utrecht, Netherlands.

Hauser, M. D. (1996). *The Evolution of Communication.* MIT Press, Cambridge, MA.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*(3), 197–243.

Hirshleifer, J. (1982). Evolutionary models in economics and law: Cooperation versus conflict strategies. *Research in Law and Economics, 4*, 1–60.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science, USA, 79*(8), 2554–2558.

Ide, N., & Veronis, J. (1998). Word sense diambiguation: The state of the art. *Computational Linguistics, 24*(1), 1–40.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press, Cambridge, England.

Kilgarriff, A., & Palmer, M. (2000). Introduction to the special issue on SENSEVAL. *Computers in the Humanities, 34*(1–2), 1–13.

Kowalski, G. (1998). *Information Retrieval Systems: Theory and Implementation.* Kluwer, Boston.

Krovetz, R., & Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems, 10*, 115–141.

Lankhorst, M. M. (1995). Automatic word categorization with genetic algorithms. In Eiben, A., Manderick, B., & Ruttkay, Z. (Eds.), *Proceedings of the ECAI'94 Workshop on Applied Genetic and other Evolutionary Algorithms* Amsterdam. Springer Verlag.

Losee, R. M. (1988). Parameter estimation for probabilistic document retrieval models. *Journal of the American Society for Information Science, 39*(1), 8–16.

Losee, R. M. (1994). Term dependence: Truncating the Bahadur Lazarsfeld expansion. *Information Processing and Management, 30*(2), 293–303.

Losee, R. M. (1996a). Learning syntactic rules and tags with genetic algorithms for information retrieval and filtering: An empirical basis for grammatical rules. *Information Processing and Management, 32*(2), 185–197.

Losee, R. M. (1996b). Text windows and phrases differing by discipline, location in document, and syntactic structure. *Information Processing and Management, 32*(6), 747–767.

Losee, R. M. (1998). *Text Retrieval and Filtering: Analytic Models of Performance.* Kluwer, Boston.

Losee, R. M. (1999). Communication defined as complementary informative processes. *Journal of Information, Communication, and Library Science, 5*(3), 1–15.

Losee, R. M., & Paris, L. A. H. (1999). Measuring search engine quality and query difficulty: Ranking with Target and Freestyle. *Journal of the American Society for Information Science, 50*(10), 882–889.

Lu, X. A., & Keefer, R. B. (1995). Query expansion/reduction and its impact on retrieval effectiveness. In *Overview of the Third Text REtrieval Conference (TREC-3),* pp. 231–239. National Institute of Standard and Technology, Computer Systems Laboratory, Gaithersburg, MD.

Manning, C. D., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, Mass.

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing, and information retrieval. *Journal of the ACM, 7*, 216–244.

Moon, S. B. (1993). *Enhancing Retrieval Performance of Full-Text Retrieval Systems Using Relevance Feedback.* Ph.D. thesis, U. of North Carolina, Chapel Hill, NC.

Newmeyer, F. J. (1986). *Linguistic Theory in America* (Second edition). Academic Press, New York.

Ney, H. (1997). Corpus-based statistical methods in speech and language processing. In Young, S., & Bloothooft, G. (Eds.), *Corpus-Based Methods in Language and Speech Processing*, pp. 1–26. Kluwer, Dordrecht.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics.* Edinburgh Univ. Press, Edinburgh.

Paass, G., & Kindermann, J. (1995). Bayesian query construction for neural network models. In Tesauro, G., Touretzky, D. S., & Leen, T. K. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 7, pp. 443–450. MIT Press, Cambridge, Mass.

Partee, B. H., Meulen, A. t., & Wall, R. E. (1990). *Mathematical Methods in Linguistics.* Kluwer, Dordrecht, The Netherlands.

Pratt, W., Raiffa, H., & Schlaifer, R. (1995). *Introduction to Statistical Decision Theory*. MIT Press, Cambridge, Mass.

Russo, J. E., Medvec, V. H., & Meloy, M. G. (1996). The distortion of information during decisions. *Organizational Behavior and Human Decision Processes, 66*(1), 102–110.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill, New York.

Sanderson, M. (1994). Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland*, pp. 142–151 New York. ACM Press.

Shaw, Jr., W. M., Wood, J. B., Wood, R. E., & Tibbo, H. R. (1991). The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research, 13*, 347–366.

Strzalowski, T. (1995). Natural language information retrieval. *Information Processing and Management, 31*(3), 397–417.

Sutcliffe, R. F. E., & Slater, B. E. A. (1995). Disambiguation by association as a practical method: Experiments and findings. *Journal of Quantitative Linguistics, 2*(1), 43–52.

Teugels, J. L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *J. of Multivariate Analysis, 32*(2), 256–268.

Turtle, H., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems, 9*(3), 187–222.

Tversky, A., Slovic, P., & Sattah, S. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95*(3), 371–384.

Van Rijsbergen, C. (1977). A theoretical basis for use of co-occurrence data in information retrieval. *Journal of Documentation, 33*(2), 106–119.

Van Rijsbergen, C. (1979). *Information Retrieval* (Second edition). Butterworths, London.

Wilks, Y., & Stevenson, M. (1998). The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Journal of Natural Language Engineering, 4*(2).

Wise, B. P., Perrin, B. M., Vaughan, D. S., & Yadrick, R. M. (1989). Evaluation of uncertain inference models III: The role of tuning. In Kanal, L. N., Levitt, T. S., & Lemmer, J. F. (Eds.), *Uncertainty in Artificial Intelligence 3*, pp. 55–62. North-Holland, Amsterdam.

Yngve, V. (1986). *Linguistics as a Science.* Indiana University Press.

Yu, C. T., Buckley, C., Lam, K., & Salton, G. (1983). A generalized term dependence model in information retrieval. *Information Technology: Research and Development, 2*(4), 129–154.