Is 1 Noun Worth 2 Adjectives? Measuring Relative Feature Utility

Information Processing & Management,
Volume 42, Number 5, 2006, Pages 1248–1259.

Robert M. Losee
CB#3360
University of North Carolina-Chapel Hill
Chapel Hill, NC 27599-3360

email: *losee at unc period edu*

April 24, 2006

Abstract

Are two adjectives worth the same as a single noun when documents are ordered based on decreasing topicality? We propose an easy to interpret single number Relative Feature Utility (RFU) measure of the relative worth of using specific linguistic or non-linguistic features or sets of features in computational systems that order or filter media, such as information retrieval and classification systems. This measure allows one to make easily interpreted claims about the relative utility of features such as parts-of-speech, term suffixes, phrases vs. single terms, annotations, hyperlinks, citations, index terms, and metadata when ordering natural language text or other media. Data is provided for the RFU for stemming characteristics, part-of-speech tags, and phrase lengths, as well as retrieval characteristics and procedures. Using this linear measure of the relative utility of features makes available a wide range of cost-benefit analyses and decision theoretic techniques, allowing the study of whether or not to use many different kinds of representational information or tagging systems, and for the design of indexing and metadata systems. Some characteristics of natural languages used in the spectrum from softer to harder sciences, as well as medical terminology, are studied.

1 Introduction

Information systems that order documents rarely incorporate all the information that is available about the documents during the sorting process. When implementing information systems, decisions are made about which linguistic components and media characteristics should be incorporated into the system and which may be omitted. The purpose of this work is to develop a tool that can be used to determine the relative utility of different linguistic and non-linguistic components when ordering media, allowing system designers and managers to select for inclusion those features that have the greatest expected effectiveness. For example, on the average, is a single noun worth the same as two adjectives, or is a hyperlink worth the same as a 3 term index phrase when documents are ordered. We refer here to the information that the ordering systems need to most efficiently sort documents, given quality, time, space, and other constraints; we are not referring to the information that the user wishes to receive.

Features that may be used for ordering documents may be categorized as linguistic or non-linguistic. Statements in a natural language are composed of a variety of features, on several different levels. While it may be desirable to understand language as a whole (Aronoff & Fudeman, 2005, p. 11), processing of natural language at the present time requires the processing of text as a set of parts, since scholars remain unable to fully understand many aspects of natural language. Understanding the relative merits of these different features of natural language is essential to deciding which features to address using computational linguistics when developing document ordering systems. A measure is provided that easily allows us to determine the relative merits of incorporating specific linguistic features into natural language based systems.

Feature topicality in systems using media ordering can exist due to a specific meaning attached to a single term or small number of terms functioning as a phrase (Church & Hanks, 1990; Fagan, 1989; Schlobinski & Schutze-Coburn, 1992). Similarly, larger grammatical units, such as sentences or paragraphs, or entire documents, may represent topic and topical-comments (Davison, 1984; Greisdorf & O'Connor, 2003; Jacobs, 2001; Shi, 2000). The topic may also be inferred from individual terms, phrases, or larger text fragments, which provide a context for inferring topicality. Individual terms have different meanings or senses associated with the various parts of the term, e.g., *psychiatry*: (*psyche* = mind, *iatreia* = healing), or the common root *run* for *run*, *runs*, *running*, and *ran*. Non-linguistic features that may be used in document ordering include characteristics such as citations and hyperlinks, the presence of specific images on a book's cover, or whether an index occurs in a book. What is the relative degree to which these linguistic and non-linguistic characteristics represent topicality

and predict the utility of a document to a system user? Term roots may carry one or more meanings or topics, and the addition of contextual or supporting information, such as suffixes, part-of-speech tags, and larger contexts can contribute to the topicality, therefore improving document ordering (Bossong, 1989; Clement & Sharp, 2003; Losee, 2001)

How does one measure how many of one feature or type of feature is equivalent in ordering power to another chosen feature or type of feature? The Relative Feature Utility may be used to empirically analyze the ordering effects of term stemming, the length of natural language phrases, the effect of using different part-of-speech labels, and various information retrieval or filtering assumptions.

2 Ordering Performance as a Utility Measure

The performance of systems incorporating characteristics may be studied by measuring how documents are ordered given different sets of characteristics. How does the ordering performance vary when incorporating just nouns or just adjectives? How would one word queries perform when they are only nouns or only adjectives? Determining the relative performance of two different characteristics used in ordering allows one to make decisions about which types of characteristics should be incorporated into a specific system. While ordering performance measures take many forms (Losee, 2000), e.g. precision, recall, average search length, or any of a number of measures of ordering performance, a linear measure that can be directly interpreted as *an occurrence of characteristic X produces performance at the same level as n occurrences of characteristic Y* can be particularly valuable when making decisions about whether characteristic X or characteristic Y should be used, along with their associated costs and benefits.

Precision is the probability that a document in the set of retrieved documents is relevant and is used as an effectiveness measure. While precision may be computed and reported over a range of recall levels, where *recall* is defined as the percent of relevant documents that have been retrieved, precision is commonly used as a measure of the quality of the retrieved set at a specific point in the ordered list of documents retrieved from a search engine. Search engines that produce initial output screens of 10 documents might have the precision computed after 10 documents have been retrieved, denoted as P_{10}.

The goal of this work is different than the goal of those who developed measures such as precision and recall. This work presents a means by which one can make a clear statement that one characteristic, for example, is two or five times as useful as another feature. The Relative Feature Utility is computed based on an analytic model of ordering performance that begins with

the Average Search Length (ASL). The ASL is the average position of relevant documents in an ordered list of documents, with the positions for N documents ranging from 1 to N, with an ASL of 1 being a single relevant document at the front of the ordered list of documents, and an ASL of N being the worst, with a single relevant document occurring at the end of the ordered list of documents. The discussion below is for single term queries. For N documents, each document with or without the single query term,

$$ASL \approx N\left[Q\mathcal{A} + (1-Q)(1-\mathcal{A})\right] + \frac{1}{2}, \tag{1}$$

where the Q value represents the probability that the ordering function is optimal, N is the number of documents in the database, and $\mathcal{A}$ is a normed average search length which scales from 0 to 1, with 0 being best-case performance and 1 being worst-case performance. The performance measure $\mathcal{A}$ is the probability that a randomly selected document is located before the average position of relevant documents and is computed as $\mathcal{A} = (1 - p + t)/2$, where p is the probability that a relevant document has the feature in question and t is the unconditional probability that a document has the feature in question. The non-approximating forms of this equation may be used for small document sets (Losee, 1998). The $[Q\mathcal{A} + (1-Q)(1-\mathcal{A})]$ component has a value between 0 and 1, and the value is converted to the Average Search Length by Equation 1. This average normalized position of a relevant document after ordering occurs is referred to as the Normalized Average Search Length (NASL):

$$NASL = Q\mathcal{A} + (1-Q)(1-\mathcal{A}).$$

The core of this equation may be computed by first measuring the ASL empirically and then algebraically solving for other variables, such as $\mathcal{Q}$ and $\mathcal{A}$. The availability of the probabilistic interpretation of $\mathcal{A}$ allows us to compute the ASL based upon parameters, rather than just upon experimental data, and to develop analytic models of ordering performance (Losee, 1998).

To illustrate the computation of NASL, consider the situation where Q is understood to be 1 (less than perfect rankings are discussed in Losee (1998) and Losee and Paris (1999)) and $\mathcal{A}$ is computed from the fact that half of the documents have the query term and three quarters of the relevant documents have the query term. In this case $\mathcal{A} = (1 - 3/4 + 1/2)/2 = 3/8$. NASL is then computed as $1 \times 3/8 + (1-1)(1-3/8) = 3/8$. This is clearly better than a random NASL of $1/2$ but is only $1/4$ of the distance from random to perfection ($NASL = 0$). ASL is similarly computed, with the primary difference being the multiplication by the number of documents. In the case of 1000 documents, for example, ASL would be $1000 \times .375 = 375$, that is, the expected position of a

relevant document would be at the 375th ordered document.

3 Relationships Between Normalized ASL Measures: Two of These Perform as Well as One of Those

Can the discussion of the relative merits of a linguistic characteristic be simplified so that one could say, for example, that two part-of-speech tags are "worth" the same as a single term suffix or a single hyperlink, or that two adjectives are worth as much as a single noun when trying to identify topicality and topic-based relevance? One can relate probabilistic distributions through use of the Kullback-Leibler information gain measure (Kullback, 1959; Losee, 1990), which examines the similarity between two probabilistic distributions. Log-odds discrimination measures (Bishop, 1995; Duda, Hart, & Stork, 2001) may also be used to measure the number of bits of information that are added by using features in a discrimination task. The ordering based Relative Feature Utility measure is somewhat different than these, given its focus on ordering, its linearity, and its ease of interpretation.

The probability that a document in the top half of the ordered list of documents is ranked ahead of the average position of the relevant documents is denoted as W. This can be obtained by multiplying $\mathcal{A}$ by 2 and then truncating anything over 1, thus $W = \min(2\mathcal{A}, 1)$. When the query term is a positive discriminator, $\mathcal{A}$ will always be less than $1/2$ and $W = 2\mathcal{A}$.

Given probability W, one may compute the number of independent occurrences of type 1 that produce the same performance as the number of independent occurrences of type 2, using $W_1 = W_2^M$. Here, M, the Relative Feature Utility (RFU), is the multiplier representing the number of occurrences of type 2 that produce the same ordering performance as a type 1 occurrence, e.g., how many adjectives produce the same ordering performance (e.g., $\mathcal{A}$) as a single noun? One can solve for M as

$$M = \frac{\log(W_1)}{\log(W_2)}. \tag{2}$$

Consider a situation where a test value has $\mathcal{A}_1 = .35$ (thus $W_1 = .7$) and a base value has the value $\mathcal{A}_2 = .45$ (thus $W_2 = .9$). The Relative Feature Utility is computed as $M = \log(.7)/\log(.9) = 3.39$. This may be interpreted as 3.39 occurrences of a type 2 event producing the same level of ordering performance as an occurrence of a type 1 event. This may be used to compare anything that contributes to ordering, such as term roots, term stems, part-of-speech tags, citations, and annotations, as well as the characteristics of phrases, documents, or entire libraries or distributed systems. The RFU may be used to

study relationships between feature characteristics, such as the number of subject headings assigned to documents compared to the length and information content of individual subject headings (Losee, 2004).

Using the Relative Feature Utility allows one to make claims that 3 of feature x produces the same performance as 1 occurrence of feature y. System designers may make choices given this knowledge, considering how expensive in terms of cost, system speed, and storage, of whether to use features x or y or both. Knowing that the precision or recall or average search length is doubled or halved when using a specific type of feature provides ordinal information that performance has improved or decreased, but the magnitude of change may be difficult to use in practical situations. However, given the knowledge that 3 of something produces the same performance as 2 of something else provides one with specific information that can be used in calculating trade-offs and decision making. For example, one might decide to use something that performs twice as well if it took less than twice the processing time. While some measures can be shown to be metrics under some circumstances (Shaw, 1986), the work here produces a linear measure (i.e., ratio data) with all the accompanying simplicity missing from non-linear measures.

Using this technique, both *types* and *tokens*, general categories and specific instances, can be analyzed. A claim could be made that using a specific occurrence of a term increases ordering performance more than 3 occurrences of a different term type or of 3 specific terms x, y, and z. Similarly, one may choose to work with entire classes, such as all nouns or all hyperlinks.

When using M as a measure of Relative Feature Utility, one may be faced with determining when a result should be treated as significant. While it has been pointed out that many information retrieval applications do not meet the assumptions made by many commonly used statistical significance tests (Van Rijsbergen, 1979), the significance of studies has been studied through the use of the t-test, sign, and Wilcoxon tests (D. Hull, 1993; Sanderson & Zobel, 2005), as well as through the use of the Kolmogorov Smirnov test (Moon, 1993). Other studies have suggested that a better focus is not only on whether two results are significantly different, but instead one should consider whether the results are significantly better than one would expect through random ordering (Shaw, Burgin, & Howell, 1997b). More *ad hoc* rules have suggested that for precision, a difference of 5% may be considered significant and a difference of 10% in precision may be considered very significant (Sparck Jones & Bates, 1977). Following this model, one may suggest that improvements in M of 5% (for a feature being used as compared to not being used) with a large database should be considered significant, that is, an M value of 1.05 should be considered significant and important, although this is rather arbitrary and is based primarily on experience with the data described in this paper. We note that

for the purposes of this study, which attempts to describe the Relative Feature Utility and illustrate its application, the numbers provided here are meant to show roughly the range of numbers that might be expected using this form of analysis.

4 The Nyltiac System and Databases

Tests were conducted to help us understand this Relative Feature Utility model and to allow us to make empirical observations. The Nyltiac system (http://Nyltiac.com), a retrieval system on the web that runs in a user's browser software, was used for most of the document ordering tasks and data manipulation tasks. Nyltiac has been used as a classroom instruction tool for several semesters and it is assumed that most of the "bugs" in major routines have been eliminated. The research version contains routines that are not available to students, while the student version that is available on the web site only contains routines that the author routinely uses in an introductory Information Retrieval course.

There are a number of ranking algorithms that could be used in ordering terms or documents. CLM (Coordination Level Matching) ranks documents by the number of term types in a document that are also in the query. For this study, the CLMF (Coordination Level Matching-Frequency) method is used, which is similar to the popular TF-IDF measure that is widely used by search engine term weightings except that CLMF treats all terms as having the same weight, avoiding the introduction of bias toward or against certain terms or classes of terms based on their relative frequency of occurrence. Terms, phrases, and documents are assigned a weight computed from the term frequency for terms that occur both in the document surrogate and in the query. In the case of single term queries, the document orderings consistent with CLMF and the TF-IDF rankings are identical.

Given a query $\{x, y\}$ and two documents A: $\{v, x, x, y\}$ and B: $\{v, y\}$, we find that the CLM value for A is 2 (1 "point" each for the presence of x and y) and for B the CLM value is 1 (for y). As an example of CLMF, using the query and documents above, the weights for A become 3 (for x, x, and y) and for B the CLMF remains 1 (for y).

In many circumstances, one may wish to analyze the effects of using a single term or a variant of a term. For many of the analyses below, the queries are broken up into individual terms or into small sets of one, two, three, or four sequential terms extracted from the original queries. In these cases, the relevance judgments for the query-document-relevance *triples* are derived from the full queries. When using these triples, where the queries are each single terms extracted from the original natural language text query, these queries are referred to as *single term queries*. There will be as many single term queries

as there were original terms in all the original natural language queries. A query with three terms with n documents would originally have n triples; if the three term query were broken up into individual single term queries, each of which would have its own triples, $3n$ triples would exist, n from each single term query. If a database had an average of 10 terms per query, 82 documents, and 35 natural language queries, using this single term model would produce $35 \times 10 = 350$ single term queries, and this, times 82 triples for each query, would produce $350 \times 82 = 28700$ single term triples.

Three databases were used for this study. The ADI463 database is the smallest database, with 463 single term queries and 82 document abstracts, representing papers presented at an American Documentation Institute Conference (Salton & Lesk, 1968). The database from which this is derived, the ADI database, is an older database and has been often used in retrieval experiments. The ADI463 database has 37996 query-document relevance judgments (triples). The second database, the MED559 database, contains 559 single term queries and 1033 documents, and is derived from the MED1033 database (Kwok, 1990; Shaw, Burgin, & Howell, 1997a). The MED559 database has 577477 query-document relevance judgments (triples). The third database, the CF683 database, is the largest, with 683 single term queries, and is derived from the first 50 queries from the CF database. All 1239 available documents are used (Shaw, Wood, Wood, & Tibbo, 1991). The CF683 database has 846237 query-document relevance judgments (triples). All these test databases have relevance judgments, normally based on a notion of topical relevance (Greisdorf & O'Connor, 2003), providing binary relevance relationships between each document and query (Baeza-Yates & Ribeiro-Neto, 1999).

We believe that having tens of thousands of query-document relevance judgments for one database and hundreds of thousands of query-document relevance judgments for the others is adequate for illustrating situations in which the Relative Feature Utility model may be applied. For example, having several hundred single term queries and hundreds of thousands of query-relevance judgments concerning nouns provides a useful snapshot of the underlying parameters of nouns in natural language. If this study were examining far less common linguistic expressions, then using larger databases would likely be necessary, but because of the large number of nouns, adjectives, and verbs that exist in each of the databases, we feel that general trends can be observed that are roughly generalizable. If one wished to gain a much more precise picture of the parameters of natural language, numerous databases representing the different contexts and sublanguages being used among humans (e.g., children's speech, children's writing, parents' speech to children, parents' speech to spouses, etc.) would probably be more valuable than using large single-domain databases, such as some of the popular large databases that are

based on documents licensed from either a single or a very small number of organizations.

Using these databases also allows us to suggest possible differences between different scientific disciplines and their terminology (Latour & Woolgar, 1986; Lodahl & Gordon, 1972; Losee, 1995; Pierce, 1992). The ADI463 database may be understood as a social science database, with more ambiguous and less precise terminology than one would find in the harder sciences, as exemplified by the MED559 and CF683 databases. The MED559 database has an additional difference in that the original natural language "queries" are statements of topic, phrases indicating what is desired. This is different than the true English language questions that serve as the basis for the single term queries in the CF683 and ADI463 databases.

4.1 Linguistic Processing

Suffixes were removed on some occasions from query and document text with the Porter stemmer (Porter, 1980). A widely used stemmer, the Porter algorithm has been used to remove suffixes in a number of languages. Stemmers may be used to find the basic morphological units representing concepts in natural language (Aronoff & Fudeman, 2005; B. Fox & Fox, 2002; Kurz & Stoffel, 2002). The use of the Porter stemmer is meant as an example of a popular stemmer, and is not meant to imply that this is the best stemmer for all situations.

Stemming algorithms bring together similar forms of terms (Cleverdon, 1967; Salton & Lesk, 1968) and have been analyzed using a number of techniques. Most frequently a retrieval performance measure is used to evaluate the relative performance of different stemming methods (D. A. Hull, 1996). This is somewhat similar to the measures of document ordering as an indicator of algorithm and feature quality. Other methods may measure the degree of stemmed term assignment to manually produced groups or meanings (Paice, 1996; Savoy, 1993), considering qualitatively cases where too much of the core term is removed, as well as cases where too little of the suffix is removed.

Terms were part-of-speech (POS) tagged through use of the Brill Part-of-Speech tagger (Brill, 1994). Although it does not provide perfect tagging, it is an established and well-understood tagger. The parts-of-speech produced by the Brill Tagger are consistent with the University of Pennsylvania Treebank. Examples of the part-of-speech tags (compared to other tag sets) are provided in (Manning & Schutze, 1999, p. 141-2). Tagger performance may be measured by direct comparison with accurate manually assigned POS tags and by the algorithm's relative speed (Murata, Ma, & Isahara, 2002; Padro & Marquez, 1998), as well as by using retrieval performance.

The numbers of unique terms in queries varies from text with part-of-speech

tags to text without such tags. For example, there are 204 unique query tokens in the ADI463 queries when untagged, but when POS tags are added, there are 231 unique query tokens. Terms such as *automated* and *coding* occur only once in the list of unique untagged tokens but can each be part-of-speech tagged in two different ways and these terms thus occur twice in the list of unique query tokens, once with each of two POS tags.

Ordering performance is computed by Nyltiac in terms of $\mathcal{A}$ values, as well as a number of other performance characteristics. Using the $\mathcal{A}$ values, the M values for empirical data may be computed.

5 Term Stems and Suffixes

Individual terms, their components, and part-of-speech tags provide information that contributes to the ordering of documents. By studying the ordering performance using individual terms, unmodified or modified, with suffixes and part-of-speech tags, one can determine the relative contribution of different natural language processing options or combinations. Table 1 shows the performance results for a system using different term options when ordering documents in the three databases.

Term roots represent the most basic semantic unit in many instances, although those who know the etymology of terms may recognize syllables in terms as carrying specific meanings from their language of origin. When language is produced, prefixes and suffixes are often added to represent features such as tense, gender, person, number, etc. Native speakers of English are aware (consciously or unconsciously) that the suffix *able* (e.g., *break, breakable*) represents an adjective, *tion* and *sion* (e.g. *explode, explosion*) represents a noun, *ly* (e.g., *happy, happily*) represents an adverb, and that the suffix *s* can represents several things, such as person (e.g., *I walk, she walks*) or plurality (e.g., *toy, toys*). The roots may represent aspects of a topic, and matching the occurrences of these roots in queries and documents should produce better than random document ordering.

The Porter algorithm for stemming (Porter, 1980) provides an effective method of isolating stems. The algorithm moves through 5 steps in the removal of suffixes. Porter tested the algorithm on 10,000 terms and found that roughly $1/3$ were reduced by Step 1. After Step 4 has been applied, roughly $2/3$ of the terms had been reduced.

Table 1 contains data showing that the performance improves as more stemming occurs. Interestingly, removing stems beyond Porter level 1 makes a great deal of difference for the ADI463 database but little for the other two databases. As the database size increases and as one moves into the harder sciences, the relative utility of stemming appears to decrease. This may be

Table 1: Performance with varying degrees of stemming. Single term queries are used. Performance with full terms and no stemming is first given, followed by results with only the first part of the Porter stemming algorithm used, then parts 1 through 4 inclusive being used, and then full Porter stemming.

Description	*ADI463*			*MED559*			*CF683*		
	$\mathcal{A}$	n	M	$\mathcal{A}$	n	M	$\mathcal{A}$	n	M
Full Terms	0.4715	463	1	0.4379	559	1	0.447	683	1
Porter 1 removed	0.4675	463	1.15	0.4317	559	1.11	0.4431	683	1.08
Porter 1-4 removed	0.4593	463	1.45	0.4305	559	1.13	0.4402	683	1.14
Full Stemming	0.4595	463	1.44	0.4288	559	1.16	0.4398	683	1.14

due to the much better initial performance achieved with no stemming for the MED559 and CF683 databases, probably due to the superior topic carrying ability of the precise terminology in the sciences (e.g., MED559 and CF683 databases) as compared to the more ambiguous terminology in the social sciences (e.g., the ADI463 database).

Analyzing the Relative Feature Utility (RFU) for the extremes in this table may provide insights. One may examine the *within-database* changes by looking at the M values, noting their general improvement as one moves down the table. Examining *between-database* M values allows the characteristics of the databases to be further studied and understood. The M value for MED559 (compared to ADI463) using Equation 2 is $\log(2\times0.4379)/\log(2\times0.4715) = 2.26$ for full terms and $\log(2\times0.4288)/\log(2\times0.4595) = 1.82$ for full stemming, with the comparable values for CF683 compared to ADI463 being 1.91 and 1.52. Clearly, the baseline, full term performance is better for the harder sciences, with the topical statements in MED559 having better performance than the questions in CF683. When full stemming is used, the harder sciences still perform better than the social sciences, but the degree of increase (RFU) is lower than the increase for full terms as one moves from softer to harder sciences.

6 The Utility of Part-of-Speech Knowledge

The above techniques also may be used to analyze the Relative Feature Utility of various parts-of-speech. This work considers terms as tagged consistent with the University of Pennsylvania Treebank, using terms tagged by the Brill Tagger with the presence of the letters *VB* in a tag as representing a verb, *NN* as

indicating a noun, *RB* as indicating an adverb, and *JJ* as indicating an adjective. In some analyses, terms labeled as nouns or adjectives are lumped together in an attempt to capture noun phrases, which may be the primary linguistic structure that captures topicality.

Part-of-speech tags, attached to a specific term, provides valuable information that can help disambiguate terms, allowing a system to determine which meaning of a term is represented by the term's presence (E. A. Fox, Nutter, Ahlswede, Evens, & Markowitz, 1988; Justeson & Katz, 1995; Losee, 2001; Rittman et al., 2004; Wilks & Stevenson, 1998). When the term *bank* occurs in a document and a query, having the term labeled as a noun in the query and meaning a river bank and labeled as a verb in the document meaning to bank a plane, the POS tagging allows the non-match to exclude documents with the term whose sense doesn't match the query sense. This is an imperfect form of disambiguation, as queries addressing financial matters (e.g., *bank* as a noun) will match with documents containing *bank* as a noun that address the inward tilt of an airplane's path.

An examination of Table 2 suggests that nouns and adjectives contributed the most to ordering tasks of those terms and parts-of-speech tested. This is consistent with the notion that nouns and noun phrases represent concrete and abstract "things" that might be the subject of a query and thus have the best potential to contribute to discriminating between relevant and non-relevant terms. Verbs and adverbs consistently assist little in the ordering process. Stemming, as shown in Tables 1 and 2, contributes a small amount to ordering performance. It is clear, at the same time, that stemming does not provide as much of an ordering improvement as does part-of-speech tagging.

In the top half of Table 2, nouns (unstemmed) have M values of approximately 2 (2.26, 1.56, and 1.88) when compared to all the parts-of-speech, that is, a noun has approximately twice the ordering performance of a randomly selected term of any part-of-speech. One finds a similar strength for nouns in the bottom half of Table 2, showing stemmed terms. The M value comparing the improvement made when changing from individual adjectives to individual nouns has the set of M values for the three databases of $\log(2 \times 0.4407)/\log(2 \times 0.4631) = 1.65$, $\log(2 \times 0.4186)/\log(2 \times 0.3683) = 0.58$, and $\log(2 \times 0.4071)/\log(2 \times 0.4302) = 1.37$ for stemmed terms. These values were computed from the $\mathcal{A}$ values in Table 2. For unstemmed terms, the M values, similarly computed, are 1.72, 0.58, and 1.28. Clearly nouns are more effective than all terms at ordering for ADI463 and for CF683, with adjectives contributing more to ordering with MED559.

The MED559 database is based on queries which are not English language questions but are instead phrases indicating the topic of the query. In these rich statements, adjectives are not merely supportive but in fact are main topic

Table 2: Terms of different parts-of-speech (POS) and corresponding $\mathcal{A}$ and M values for ADI463, MED559, and CF683 databases. The top half of the table shows values with full terms (unstemmed) for all terms of the given part-of-speech, while the bottom half of the Table shows values for stemmed terms (term roots). Single term queries are used. Note that the n values may exceed the number of single query terms because the POS tagged query features include punctuation marks, as well as traditional natural language terms.

Description	*ADI463*			*MED559*			*CF683*		
	$\mathcal{A}$	n	M	$\mathcal{A}$	n	M	$\mathcal{A}$	n	M
Full Terms:									
All POS	0.4737	531	1	0.4467	684	1	0.449	687	1
Verbs	0.4764	63	0.90	0.4767	34	0.42	0.4975	84	0.05
Adverbs	0.4904	15	0.36	0.4865	14	0.24	0.4911	7	0.17
Adjectives	0.4657	47	1.32	0.3699	64	2.67	0.427	73	1.47
Nouns	0.4425	206	2.26	0.4194	279	1.56	0.4085	256	1.88
Nouns & Adjectives	0.4468	253	2.08	0.4101	343	1.76	0.4126	329	1.79
Stems removed:									
Verbs	0.4763	63	0.90	0.4767	34	0.42	0.4975	84	0.05
Adverbs	0.4904	15	0.36	0.4865	14	0.24	0.4911	7	0.17
Adjectives	0.4631	47	1.42	0.3683	64	2.71	0.4302	73	1.40
Nouns	0.4407	206	2.34	0.4186	279	1.58	0.4071	256	1.91
Nouns & Adjectives	0.4449	253	2.16	0.4092	343	1.78	0.4124	329	1.79

carriers. The first 10 adjectival occurrences in the CF683 database queries are *physical, submucosal, respiratory, salivary, different, normal, respiratory, abnormal, other,* and *therapeutic*, while the first 10 adjectival occurrences in the MED559 database queries are *crystalline, cerebrospinal, partial, bronchial, fatty, placental, normal, fatty, ventricular* and *septal*. The adjectives in MED559 are more subject-bearing than adjectives in CF683, leading to the significant difference between the M values for MED559 and the other databases.

Comparing nouns and verbs shows that nouns contribute much more to ordering than do verbs. The M values for unstemmed nouns (improvement over unstemmed verb performance) is $\log(2 \times 0.4425)/\log(2 \times 0.4764) = 2.53$, $\log(2 \times 0.4194)/\log(2 \times 0.4767) = 3.68$, and $\log(2 \times 0.4085)/\log(2 \times 0.4975) = 40.32$ for the ADI463, MED559, and CF683 databases, respectively. These values were computed from the $\mathcal{A}$ values in Table 2. Nouns carry more topicality for ordering in the harder sciences than in the social sciences, and thus the M values are larger for the MED559 and CF683 databases than for the ADI463 database. When comparing the MED559 and CF683 databases, the MED559 database expresses much of its topicality through its adjectives, with the nouns in the CF683 database carrying a much greater degree of topicality than is found in the MED559 database. Nouns seem to be better indicators of topicality in the harder sciences than in the social sciences.

Table 3: Phrase sizes and corresponding $\mathcal{A}$ and M values for ADI463, MED559, and CF683 databases. M values are computed from the row where $M = 1$ above the value, with two columns of M values shown for each database: a column with $M = 1$ computed for full terms of size 1 and the second M column having $M = 1$ for each type of phrase of length 1. The upper third of the Table contains no part-of-speech information and full terms, the middle third contains only full nouns and adjectives, and the last third contains only stemmed nouns and adjectives.

Phrase Size	*ADI463*				*MED559*				*CF683*			
	$\mathcal{A}$	n	M		$\mathcal{A}$	n	M		$\mathcal{A}$	n	M	
Full Terms:												
1	0.4715	463	1	1	0.4379	559	1	1	0.447	683	1	1
2	0.4622	428	1.34	1.34	0.4239	529	1.25	1.25	0.4326	633	1.29	1.29
3	0.4494	393	1.82	1.82	0.4173	501	1.36	1.36	0.4252	583	1.45	1.45
4	0.4367	358	2.31	2.31	0.4126	473	1.45	1.45	0.4181	533	1.60	1.60
Nouns & Adjectives:												
1	0.4468	253	1	1.92	0.4101	343	1	1.49	0.4126	329	1	1.71
2	0.4029	218	1.92	3.68	0.356	313	1.71	2.56	0.3588	279	1.73	2.96
3	0.3691	183	2.70	5.17	0.3199	283	2.25	3.37	0.3234	230	2.27	3.89
4	0.334	148	3.59	6.87	0.2972	255	2.62	3.92	0.2955	182	2.74	4.69
Stemmed Nouns & Adjectives:												
1	0.4449	253	1	1.99	0.4092	343	1	1.51	0.4122	329	1	1.72
2	0.4017	218	1.87	3.73	0.3545	313	1.72	2.59	0.3575	279	1.74	2.99
3	0.3676	183	2.63	5.24	0.3179	283	2.26	3.41	0.3204	230	2.30	3.97
4	0.3313	148	3.53	7.01	0.2946	255	2.64	3.99	0.2918	182	2.79	4.80

7 Phrases and Groups of Terms

Queries and document terms were treated above as single term queries. In this section, sequences of n terms are considered as *phrases* of size n. Given a phrase, this work matches by using the bag-of-words model, where terms matching between query and document are matched in any order. Thus, $\{a, b, c\}$ would perfectly match with $\{c, b, a\}$ and $\{a, c, b\}$. Each possible phrase of size n that can be derived from the queries as a set of successive terms is used to produce new query-document-relevance relationships. With phrases of size 2, there are 3 phrases from the original query $\{a, b, c, d\}$: the new queries $\{a, b\}$, $\{b, c\}$, and $\{c, d\}$. The number of queries for a given database in Table 3 decreases as the phrase size increases because there is usually 1 fewer phrase of size n in a given query than there is of size $n - 1$.

The bag-of-words model is used here rather than trying to match sequences because it was found that as sequences increase in length, very few exact matches occur. Using the full bag-of-words model with sequential phrases results in much better full and partial matching, thus capturing the topicality inherent in phrases (Haas & Losee, 1994; Jacquemin, 1996; Losee, 1994a, 1996).

The data shown in Table 3 provides two sets of M values for each database, representing different starting points for the computation of the base value for M. The data shows that larger sets of terms contribute more to ordering performance than do smaller sets (Losee, 2004). As more features are made available, more information is available to allow us to discriminate between documents of different topical-relevance categories.

One can approximate the study of phrases that may be thought of as noun phrases by examining phrases containing only nouns and adjectives. Ordering with only nouns and adjectives performs better than with the full untagged terms. Table 3 shows comparable phrase sizes being superior when they are noun phrases (nouns and adjectives) compared to using no part-of-speech tags. Stemming adds little beyond what is already obtained with nouns and adjectives, suggesting that someone designing a system might gain significantly by limiting phrases to noun phrases but would gain much less by stemming.

Beginning with the top of Table 3, data suggests that ordering for MED559 and CF683 databases is superior to performance with ADI463 databases. The M values for MED559 and CF683 databases for single full term phrases, compared to an $M = 1$ for the ADI463 database, may be computed as 2.26 and 1.91, respectively. Considering the 4 term phrases with only nouns and adjectives, data about MED559 and CF683 suggests that the M values, based on the ADI463 values, may be computed as 1.42 and 1.32. respectively, and for stemmed nouns and adjectives are 1.29 and 1.31.

This decrease in M values is probably due in part to the much stronger

topic-carrying ability of the full terms in the harder sciences. The social science database carries less topicality initially and thus has greater room for improvement (and does improve more) as one adds part-of-speech information and stemming.

8 Large Groups of Terms: Documents

One can extend the power of phrases to use all terms in a query when matching queries and documents for retrieval purposes. Table 4 shows results for more traditional information retrieval experimentation. The M values for MED559 and CF683 clearly take a leap when scanning down the table when one moves to stemmed nouns, which clearly outperform other modifications or options that are used. The upper-bounds values show another leap, as one would expect as one moves from empirical data to upper limits. Note that the upper bounds ranking is determined by ranking documents by their retrospective expected precision values (Losee, 1994b), first retrieving the set of documents with the same characteristics, vis-à-vis the query, which has the highest precision, then the set of documents with the same characteristics as the query with the second highest precision, and so forth.

Several other weighting systems are used in this table. TF-IDF represents term frequency multiplied by the Inverse Document Frequency (IDF) weight. Because the CLMF weight used here is similar to TF-IDF but with the IDF weight held constant, the ordering performance values for CLMF and TF-IDF values are similar. The Binary Independent (BI) and Two Poisson Independent (TPI) methods (Lewis, 1998; Losee, 1988) use full knowledge (omniscience) or retrospective parameter estimation and use the initial parameters of $1, 2$ (a prior beta distribution with parameters 1, 2 that could be used with relevance feedback) for both relevant and non-relevant documents (Losee, 1988).

9 Discussion & Conclusion

When information carrying units are ordered by their probability of topical relevance, the performance shown in the ordering, and the use of features when ordering, can be measured. These values, in turn, may be compared so that a given feature may be said to be n times as helpful as another feature using the Relative Feature Utility (RFU) measure, a linear measure of relative performance. System designers and users who have the choice of which features a system should use may then make a rational, cost-effective decision about which features to include by computing the M or RFU value. Features may be ordered and the better and most cost effective features used. Clearly, a low M

Table 4: Traditional full queries for information retrieval processing.

Description	*ADI463*			*MED559*			*CF683*		
	$\mathcal{A}$	n	M	$\mathcal{A}$	n	M	$\mathcal{A}$	n	M
Full Terms & CLMF	0.3995	35	1	0.3946	30	1	0.421	50	1
Stemmed Terms & CLMF	0.3574	35	1.50	0.3799	30	1.16	0.4087	50	1.17
Full Terms & POS & CLMF	0.3991	35	1.00	0.3909	30	1.04	0.4247	50	0.95
Stemmed Terms & POS & CLMF	0.3977	35	1.02	0.3903	30	1.05	0.4251	50	0.94
Stemmed Nouns & CLMF	0.3406	35	1.71	0.1621	30	4.76	0.326	50	2.49
Stemmed Nouns & Omniscient BI	0.358	35	1.49	0.1745	30	4.45	0.3006	50	2.96
Stemmed Nouns & Omniscient TPI	0.3547	35	1.53	0.1714	30	4.52	0.3373	50	2.29
Stemmed Nouns & TF-IDF	0.3576	35	1.49	0.1776	30	4.37	0.3256	50	2.49
Stemmed Nouns & Upper Bounds	0.1727	35	4.74	0.1159	30	6.18	0.1559	50	6.78

feature that occurs frequently may be a better choice than a higher M valued feature that occurs rarely; however, the rare but high M valued feature may be the better choice in situations where there is a significant cost associated with using each feature occurrence, possibly due to storage and processing costs.

The empirical results presented above help us to understand how terms and their features have utility in the ordering of documents, as do other features in documents. Stemming terms consistently resulted in improved performance, as did the addition of part-of-speech information about nouns and adjectives. We have also examined how longer phrases resulted in better ordering performance. Table 2 shows precisely how much more topically powerful nouns are than most other parts-of-speech. Using this knowledge would allow a natural language engineer to decide how much more useful a noun would be than an adjective when considering the cost effectiveness of incorporating specific features in document ordering systems or when designing metadata systems, for example.

Are two adjectives as useful as a noun when applied to ordering tasks? This work has provided a measure that can be used to answer this form of question. More specifically, we have suggested that the answer is close to *yes* for a social science database ($M = 1.7$) but is clearly *no* with the medical databases. Further studies will need to be conducted to make more definitive statements about the relationships between different linguistic and non-linguistic features across a range of different environments.

References

Aronoff, M., & Fudeman, K. (2005). *What is morphology?* Malden, MA: Blackwell.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Harlow, England: Addison Wesley.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Bossong, G. (1989). Morphemic marking of topic and focus. *Belgian Journal of Linguistics*, *4*, 27–51.

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.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Clement, R., & Sharp, D. (2003). Ngram and Bayesian classification of documents for topic and authorship. *Literary & Linguistic Computing*, *18*(4), 423–447.

Cleverdon, C. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, *30*, 172–181.

Davison, A. (1984). Syntactic markedness and the definition of sentence topic. *Language*, *60*(4), 797–846.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

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.

Fox, B., & Fox, C. J. (2002). Efficient stemmer generation. *Information Processing and Management*, *38*, 547–558.

Fox, E. A., Nutter, J. T., Ahlswede, T., Evens, M., & Markowitz, J. (1988). Building a large thesaurus for information retrieval. In *Proceedings of the second conference on applied natural language processing* (pp. 101–108). Morristown, NJ: Assocation for Computational Linguistics.

Greisdorf, H., & O'Connor, B. (2003). Nodes of topicality: Modeling user notions of on topic documents. *Journal of the American Society for Information Science and Technology*, *54*(14), 1296–1304.

Haas, S. W., & Losee, R. M. (1994). Looking in text windows: Their size and composition. *Information Processing and Management*, *30*(5), 619–629.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval, pittsburgh, pa* (pp. 329–338). New York: ACM Press.

Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, *47*(1), 70–84.

Jacobs, J. (2001). The dimensions of topic-comment. *Linguistics*, *39*(4), 641–681.

Jacquemin, C. (1996). What is the tree that we see through the window: A linguistic approach to windowing and term variation. *Information Processing and Management*, *32*, 445–448.

Justeson, J. S., & Katz, S. M. (1995). Principled disambiguation: Discriminating adjective senses with modified nouns. *Computational Linguistics*, *21*(1), 1–27.

Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

Kurz, T., & Stoffel, K. (2002). Going beyond stemming: Creating concept signautres of complex medical terms. *Knowledge-Based Systems*, *15*, 309–313.

Kwok, K. L. (1990). Experiments with a component theory of probabilistic information retrieval based on single terms as document components. *ACM Transactions on Information Systems*, *8*(4), 363–386.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*.

Princeton, NJ: Princeton Univ. Press.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98, 10th European conference on machine learning* (pp. 4–15). Berlin: Springer.

Lodahl, J. B., & Gordon, G. (1972). The structure of scientific fields and the functioning of university graduate departments. *American Sociological Review*, *37*, 57–72.

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. (1990). *The science of information: Measurement and applications*. San Diego: Academic Press.

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

Losee, R. M. (1994b). Upper bounds for retrieval performance and their use measuring performance and generating optimal Boolean queries: Can it get any better than this? *Information Processing and Management*, *30*(2), 193–203.

Losee, R. M. (1995). The development and migration of concepts from donor to borrower disciplines: Sublanguage term use in hard & soft sciences. In M. E. D. Koenig & A. Bookstein (Eds.), *Proceedings of the fifth international conference on scientometrics and informetrics* (pp. 265–274). Medford, NJ: Learned Information.

Losee, R. M. (1996). 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*. Boston: Kluwer.

Losee, R. M. (2000). When information retrieval measures agree about the relative quality of document rankings. *Journal of the American Society for Information Science*, *51*(9), 834–840.

Losee, R. M. (2001). Natural language processing in support of decision-making: Phrases and part-of-speech tagging. *Information Processing and Management*, *37*(6), 769–787.

Losee, R. M. (2004). A performance model of the length and number of subject headings and index phrases. *Knowledge Organization*, *31*(4), 245–251.

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.

Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Moon, S. B. (1993). *Enhancing retrieval performance of full-text retrieval systems*

using relevance feedback. Unpublished doctoral dissertation, U. of North Carolina, Chapel Hill, NC.

Murata, M., Ma, Q., & Isahara, H. (2002). Comparison of three machine-learning methods for Thai part-of-speech tagging. *ACM Transactions on Asian Language Information Processing*, *1*(2), 145–158.

Padro, L., & Marquez, L. (1998). On the evaluation and comparison of taggers: The effect of noise in testing corpora. In *COLING-ACL '98, 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics* (pp. 997–1002). San Francisco: Morgan Kaufmann.

Paice, C. D. (1996). Method for evaluation of stemming algorithms based on error counting. *Journal of the American Society for Information Science*, *47*(8), 632–649.

Pierce, S. J. (1992). On the origin and meaning of bibliometric indicators: Journals in the social sciences, 1886–1985. *Journal of the American Society for Information Science*, *43*(7), 477–487.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Rittman, R., Wacholder, N., Kantor, P., Ng, K. B., Strzalkowski, T., & Sun, Y. (2004). Adjectives as indicators of subjectivity in documents. In *Proceedings of the 67th ASIST annual meeting* (Vol. 41, pp. 349–359). Providence, RI: ASIST.

Salton, G., & Lesk, M. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, *15*(1), 8–36.

Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, & N. Ziviani (Eds.), *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, salvador, brazil* (pp. 162–169). New York: ACM Press.

Savoy, J. (1993). Stemming of French words based on grammatical categories. *Journal of the American Society for Information Science*, *44*(1), 1–9.

Schlobinski, P., & Schutze-Coburn, S. (1992). On the topic of topic and topic continuity. *Linguistics*, *30*(1), 89–121.

Shaw, W. M., Jr. (1986). On the foundation of evaluation. *Journal of the American Society for Information Science*, *37*(5), 346–348.

Shaw, W. M., Jr., Burgin, R., & Howell, P. (1997a). Performance standards and evaluations in IR test collections: Cluster based retrieval models. *Information Processing and Management*, *33*(1), 1–14.

Shaw, W. M., Jr., Burgin, R., & Howell, P. (1997b). Performance standards and evaluations in IR test collections: Vector-space and other retrieval models. *Information Processing and Management*, *33*(1), 15–36.

Shaw, W. M., Jr., Wood, J. B., Wood, R. E., & Tibbo, H. R. (1991). The cystic fibro-

sis database: Content and research opportunities. *Library and Information Science Research*, *13*, 347–366.

Shi, D. (2000). Topic and topic-comment constructions in Mandarin Chinese. *Language*, *76*(2), 383–408.

Sparck Jones, K., & Bates, R. G. (1977). *Research on automatic indexing 1974-1976* (Tech. Rep.). Cambridge, England: Computer Laboratory, University of Cambridge.

Van Rijsbergen, C. (1979). *Information retrieval* (Second ed.). London: Butterworths.

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), 135–143.