Research Paper ■

Corpus-based Statistical Screening for Phrase Identification

WON KIM, PHD, W. JOHN WILBUR, MD, PHD

Abstract **Purpose:** The authors study the extraction of useful phrases from a natural language database by statistical methods. The aim is to leverage human effort by providing preprocessed phrase lists with a high percentage of useful material.

Method: The approach is to develop six different scoring methods that are based on different aspects of phrase occurrence. The emphasis here is not on lexical information or syntactic structure but rather on the statistical properties of word pairs and triples that can be obtained from a large database.

Measurements: The Unified Medical Language System (UMLS) incorporates a large list of humanly acceptable phrases in the medical field as a part of its structure. The authors use this list of phrases as a gold standard for validating their methods. A good method is one that ranks the UMLS phrases high among all phrases studied. Measurements are 11-point average precision values and precision-recall curves based on the rankings.

Result: The authors find of six different scoring methods that each proves effective in identifying UMLS quality phrases in a large subset of MEDLINE. These methods are applicable both to word pairs and word triples. All six methods are optimally combined to produce composite scoring methods that are more effective than any single method. The quality of the composite methods appears sufficient to support the automatic placement of hyperlinks in text at the site of highly ranked phrases.

Conclusion: Statistical scoring methods provide a promising approach to the extraction of useful phrases from a natural language database for the purpose of indexing or providing hyperlinks in text.

■ **J Am Med Inform Assoc.** 2000;7:499–511.

Modern computer-based retrieval systems have the potential to retrieve from a large database those documents that satisfy a Boolean query composed of virtually any words and phrases the operator may desire. Given this power, it is reasonable to ask what purpose indexing could serve. There is actually the potential for a large benefit. As shown by a number of studies,[1–6] there is great inconsistency in the terms people use to describe the same subject. In the words of Bates,[7] ''In study after study, across a wide range of environments, it has been found that for any target topic, people will use a very wide range of different terms, and no one of those terms will occur very frequently.'' Indexing can alleviate this problem by expanding the list of terms by which a document may be accessed. Thus, one path to improved indexing is to obtain a list of terms (words and phrases) sufficient to include a high percentage of the terms that people will actually use in querying a database, and add sufficient synonymy information to allow a query expressing a particular concept to access those documents that are indexed with an expression synonymous with the query.

The Unified Medical Language System (UMLS)[8] includes not only a large list of important terms but also a synonymy capability relating these terms in the Metathesaurus. It is intended, among other things, to provide a solution to the indexing problem just out-

Affiliation of the authors: National Library of Medicine, Bethesda, Maryland.

Correspondence and reprints: Won Kim, PhD, National Library of Medicine, Building 38A, Room 8S806, 8600 Rockville Pike, Bethesda, MD 20894; e-mail: ⟨wonkim@ncbi.nlm.nih.gov⟩.

Received for publication: 12/9/99; accepted for publication: 3/13/00.

lined.[8,9] The system as it stands is, of course, incomplete[10] and will, for the forseeable future, stand to benefit from increased coverage of the latest terminology in the various fields covered by MEDLINE. Our hypothesis is that statistical information about the occurrence of phrases in MEDLINE can provide a useful screen for candidate phrases that are of similar quality to the material already in the UMLS. A person generally does not possess the kind of information that is available in this way. This information can, however, be readily obtained by automatic processing and can serve as a guide to terms that would make useful additions to UMLS. Such guidance may be important, given the limited human resources that are available to integrate terminology into the Metathesaurus.

The development of a controlled vocabulary of indexing phrases is not the only use to which our methods can contribute. The ranking of phrases by quality can also be used as an aid to automatically place hyperlinks in text. We are currently involved in a project to link phrases in MEDLINE documents to appropriate sections and subsections of books in the field of biomedicine that may provide the reader with additional information about the subjects of the phrases. Here the most useful phrases in a MEDLINE record are marked as hot links that are "clickable" to reach a book or books of potential interest. The first book, *Molecular Biology of the Cell*,[11] is available at http://www.ncbi.nlm.nih.gov/entrez. To see the book links, the user must select a single document and, when it is displayed, click on the "book" link to the right of the document. In the applications sections of this paper, we show how this type of linkage can be produced automatically on the basis of phrase ranking. The hyperlinks viewable at http://www.ncbi.nlm.nih.gov/entrez differ mainly in incorporating some human review of the phrase lists used.

Several methods have been used historically in attempts to extract useful words and phrases from document collections for purposes of indexing. Since Luhn's pioneering work on indexing,[12,13] the importance of term frequency information has been recognized. The frequencies of both phrases and the words that compose them are important in the phrase extraction method of Jones et al.[14] A second kind of information that can be helpful in indexing is the distribution of frequencies of a term within documents. It has been proposed that non-content-bearing terms are well modeled by a single Poisson distribution, whereas content-bearing terms require a two-Poisson or some more complicated model.[15–17] In fact, one of our scoring methods is based on the degree that a term's distribution deviates from a Poisson distribution. The greater this deviation, the more likely that the term is a useful one. In addition to this method we employ other relatively simple scoring methods based on term frequencies, co-occurrence, and word suffixes. The objective is to locate phrases that are grammatically acceptable and specific in their meaning, yet occur with sufficient frequency in the database to make them useful additions to UMLS.

There are many methods of noun phrase extraction based on natural language processing that we have not examined. Proprietary methods such as CLARIT[18] and NPtool[19] were not of interest, since we seek to understand the methods in as much detail as possible. The transformation-based parsing developed by Brill,[20,21] hidden Markov part-of-speech tagging as in the Xerox Tagger,[22] and parsing based on a probabilistic grammer as in CHOPPER[23] are potentially of greater interest. However, these are complex tools designed for a different task than ours. They seek to assign part-of-speech tags as a basis for natural language parsing, whereas we seek to identify those phrases that are not only syntactically correct but also readily recognizable by human beings as useful and descriptive of a subject area. Even if natural language parsing methods can contribute to the accomplishment of our task, we must still ask whether their complexity is necessary to its accomplishment. We seek to show in what follows that simpler methods suffice. We will examine the Xerox Tagger to show that it adds little to what can be accomplished by our scoring methods.

Another phrase extraction task that has been studied is phrase extraction with the purpose of improving retrieval by expanded automatic indexing on test collections. The methods of phrase identification are based on part-of-speech tagging as well as some statistical methods. This area is exemplified by the work of Fagan[24] and Lewis and Croft.[25] Interestingly, while there has been some success with this approach in improving retrieval, the results are not consistently good. This led Lewis and Jones[26] to comment that "... automatically combining single indexing terms into multiword indexing phrases or more complex structures has yielded only small and inconsistent improvements over the simple use of multiple terms in a query." We mention this area mainly to distinguish it from our own work. Different goals and different methods of evaluation characterize the two approaches. Instead of seeking to improve retrieval in some automatic system, we seek to identify those phrases that are the most user friendly, and we evaluate our success by how well we are able to identify a set of phrases (UMLS) that are maintained by human beings because they are found descriptively useful.

We begin with a description of the different data sets we study and how they are constructed. We then present the scoring methods that are designed to distinguish useful phrases from simple co-locations of terms. We describe our approach to evaluation of the scoring methods and present results on the effectiveness of the scoring methods when applied to a large database of MEDLINE records. Besides the six scoring methods that we find useful, we evaluate two other methods and find that they do not add significantly to overall effectiveness. We discuss application of our methods to extraction of candidates for UMLS and also as a procedure for marking text with hyperlinks. The paper concludes with a discussion and description of future directions.

Data Sources and Preparation

We consider word pairs and word triples from two different sources. The first source is the UMLS[8,27] developed by the National Library of Medicine. The UMLS (9th edition, 1998) was obtained from the National Library of Medicine on CD. (Information regarding its availability for research purposes may be found at http://www.nlm.nih.gov/research/umls/umlsmain.html.) Our second data source is the set of 304,057 MEDLINE records with abstracts and entry dates in the year 1996. We shall refer to this document set as MED96. These two data sets are processed somewhat differently, because they differ considerably in content. However, we will use a procedure to normalize text strings that is the same for both. We normalize text in three steps: All alphabetic characters are lowercased, all non-alphanumeric characters are replaced by blanks, and multiple blank spaces between words are converted to single blank spaces.

The UMLS is processes as follows: First, all text strings are obtained from the UMLS "mrcon" (concept name) file. From the resulting set of strings, any containing punctuation marks or stop words are deleted. For this purpose a list of 310 common stop words is used. Finally, the remaining strings undergo normalization and removal of any duplicates. The result is a set we denote by U_{all}. This is the set of all phrases that we obtain from UMLS.

From U_{all} we extract the subset of strings consisting of two words each. The result is 156,086 word pairs, denoted by U_2. In the same way we extract all three-word phrases from U_{all}. The result is 103,367 word triples, denoted by U_3.

MED96 is processed somewhat differently. We first process the titles and abstracts of the MED96 records, breaking at punctuation marks and stop words. The resulting set of strings is normalized and made unique, to produce the set M_{seg}. This is the set of longest phrases that we obtain from MED96. By M_2 we denote the set of all contiguous word pairs that can be obtained from the members of M_{seg}. For example, the four-word string "escherichia coli cell growth" from M_{seg} yields the three overlapping two-word phrases "escherichia coli", "coli cell", and "cell growth" in M_2. By M_3 we denote the set of all contiguous word triples that can be obtained from the same source.

The difference in the processing of the UMLS and MED96 is perhaps worth emphasizing. The strings in U_{all} are essentially a subset of the strings that occur in the UMLS "mrcon" file, except for lower casing, and as such by and large represent syntactically reasonable and semantically meaningful phrases. The subset U_2 is just those strings in U_{all} that are composed of two words. There are longer phrases in U_{all} that could be broken up into contiguous two-word phrases and added to U_2, but we do not do this because we do not know whether these would be of high quality. The same applies to the derivation of U_3. The U_2 and U_3 sets represent our gold standard for good phrases, and we seek to keep their quality as high as possible.

On the other hand, M_{seg} is a large set of strings that are obtained from all the text in MED96. Many of these are not, as phrases, of high quality. The M_2 set is derived from M_{seg} by taking all those strings in M_{seg} that consist of two words as well as all those contiguous word pairs that may be obtained from longer phrases in M_{seg}. The longer phrases in M_{seg} are broken up in this way and added to M_2 because, even if the longer phrases are of poor quality, some two-word substrings may be of good quality and such potential should not be ignored. The same basic method applies to the derivation of M_3. It will then be the task of the scoring procedures that we introduce to separate the good from the bad.

Scoring Methods

In this section, we define the various scoring methods we want to apply to the word pairs in the set M_2 and the word triples in the set M_3 extracted from the MED96 database. Our goal is to define scoring methods that will allow us to find the most useful phrases occurring in a database. We only define the methods and give some justification for their choice here. Their systematic evaluation is the subject of the next sections. We begin by describing scoring methods for the word pairs in the set M_2. When these have been described we indicate the modifications necessary for application of the same methods to M_3.

Method I

Given a word pair in the set M_2, we perform a simple count of the number of MED96 documents that contain that word pair (*phrase frequency*). Dividing this count by the normalization factor N (the size of MED96), the corresponding score s_1 is

$$s_1 = \frac{phrase\ frequency}{N} \quad (1)$$

The normalization factor is a constant and is optional here, but it might allow one to compare results across databases more readily.

Rationale: Phrases as well as single words follow a Zipf-like distribution, with a plethora of very low frequency phrases and progressively fewer examples in the higher-frequency categories. Rare phrases are of only limited value as discriminators.[28] Naturally, the UMLS tends to avoid very low frequency terms, which explains the utility of frequency as a scoring method.

Method II

Given a word pair in the set M_2, we count the number of documents in MED96 that contain both words, even if not as a contiguous pair. The result is called the co-occurrence, and the score s_2 is

$$s_2 = \frac{phrase\ frequency}{co\text{-}occurrence} \quad (2)$$

It is evident that this score always lies between 0 and 1.

Rationale: As an example, consider two word pairs, ''diabetes mellitus'' and ''wide tumor.'' For ''diabetes mellitus,'' the phrase frequency (the number of the MED96 documents that contain ''diabetes mellitus'') is 2,465, the co-occurrence (the number of the MED96 documents that contain both ''diabetes'' and ''mellitus'' but not necessarily as a contiguous pair) is 2,468, and thus the score s_2 is 0.99. For ''wide tumor,'' the phrase frequency is 3, the co-occurrence is 352, and the score s_2 is 0.008. Two words that tend to co-occur only in the form of a phrase often form a high-quality phrase.

Method III

Given a phrase in the set M_2, we examine all occurrences of the phrase throughout the text of MED96. As described in the previous section, the text of MED96 is broken at stop words and punctuation marks, and the resulting phrases compose the elements of M_{seg}. Each occurrence of a phrase that immediately precedes one of these break points (at a stop word or a punctuation mark) is counted in $phrase_{end}$ for that phrase. For example, the word pair ''lipoprotein cholesterol'' occurs in the sentence fragment ''. . . serum total and high-density lipoprotein cholesterol, C-reactive protein, and plasma fibrinogen.'' Here it occurs just before a comma, and hence this occurrence will contribute 1 to the score $phrase_{end}$ for the phrase ''lipoprotein cholesterol.'' The meaning of ''end'' in this context is that ''lipoprotein cholesterol'' is at the right-hand end of the longer phrase ''high-density lipoprotein cholesterol'' that this sentence fragment contributes to M_{seg}. In the same sentence fragment is also the phrase ''density lipoprotein,'' but since this occurrence of ''density lipoprotein'' does not immediately precede a stop word or punctuation mark, it does not add to the score $phrase_{end}$ for ''density lipoprotein.'' Again, normalizing by the total number of MED96 documents, the score s_3 is

$$s_3 = \frac{phrase_{end}}{N} \quad (3)$$

Rationale: The scoring method s_3 is a quasi-syntactic categorization. The head of a phrase tends to occur at the right-hand end.[29] The number of times that a phrase ends at a stop word or a punctuation mark is a measure of the likelihood that its last word is a head and, therefore, of whether the phrase can stand alone. For example, ''central nervous'' will be followed immediately by a stop word or punctuation mark much less frequently than will the phrase ''nervous system.''

Method IV

The score s_4 is obtained as an odds ratio based on the last three characters of the last word in the phrase. The definition is

$$s_4 = \frac{p(good\ phrase|l_1l_2l_3)}{p(good\ phrase)} \quad (4)$$

where the number $p(good\ phrase|l_1l_2l_3)$ is the probability of being a good phrase given the last three letters $l_1l_2l_3$. From a simple rearrangement of the Bayes theorem, we can infer s_4, i.e.,

$$\frac{p(good\ phrase|l_1l_2l_3)}{p(good\ phrase)} = \frac{p(l_1l_2l_3|good\ phrase)}{p(l_1l_2l_3)} \quad (5)$$

where $p(l_1l_2l_3|good\ phrase)$ is obtained as the distribution of the last three letters of the last word over all phrases of U_{all}, and $p(l_1l_2l_3)$ is obtained as the distribution of the last three letters of the last word over all the phrases in M_2.

Rationale: The scoring method s_4 is based on the characteristic suffixes that tend to be applicable to differ-

ent word classes and different parts of speech. For example, if the last three characters of the last word in a word pair are "-ely," as in "bind cooperatively," the phrase may not be of very high quality ($s_4 = 0.044$). However, if the last three characters of the last word in a word pair are "-ine" (often the suffix of a chemical or medicine), such as "basophil histamine," "biogenic amine," and "catalytic histadine," the phrase may be of high quality ($s_4 = 2.45$).

Method V

Our next scoring method is based on the hypergeometric distribution.[30] For a given word pair in the set M_2, let n_f equal the number of MED96 documents that contain the first word in the pair and n_s equal the number of MED96 documents that contain the second word in the pair. Again, let N denote the total number of MED96 documents. If x denotes the co-occurrence of the two words and if we assume the words are randomly distributed, then x obeys the *hypergeometric probability* distribution, defined by

$$p(x) = \binom{n_f}{x}\binom{N - n_f}{n_s - x}\binom{N}{n_s}^{-1} \quad (6)$$

Using this distribution we may obtain the P value, i.e., the probability that the actual co-occurrence is as great as or greater than the observed co-occurrence if the words are assumed to be randomly distributed:

$$P \text{ value} = p(x \geq \textit{co-occurrence}) = \sum_{x=\textit{co-occurrence}}^{\min(n_f, n_s)} p(x) \quad (7)$$

where $\min(n_f, n_s)$ is the smaller of the numbers n_f and n_s. Then s_5 is given by

$$s_5 = -\log(P \text{ value}) \quad (8)$$

Rationale: If the observed co-occurrence of a word pair is quite above the expected value for a random incident, the phrase may be a useful one. For example, for the word pair "surgically curable" we have n_f = 1,583, n_s = 148, N = 304,057, and *co-occurrence* = 3. The estimated co-occurrence (E_{co}) from the hypergeometric distribution is

$$E_{co} = \frac{n_f \cdot n_s}{N} = 0.77$$

The words "surgically" and "curable" appear together at a near random level in the database. However, for the word pair "immunodeficiency virus," we have n_f = 3,505, n_s = 11,143, N = 304,057, and *co-occurrence* = 2,845. Also, the expected co-occurrence (E_{co}) from the hypergeometric distribution is

$$E_{co} = \frac{n_f \cdot n_s}{N} = 128$$

The observed *co-occurrence* (= 2,845) for the words "immunodeficiency" and "virus" in the MED96 database is far above the random level (E_{co} = 128). The score s_5, which is the negative logarithm of the P value, is the measure of the discrepancy from a random incident (s_5 = 1.36859 for the word pair "surgically curable" and s_5 = 3,527.45 for the word pair "immunodeficiency virus").

Method VI

Our final scoring method is based on the distribution of the within-document term frequencies. We define a randomly distributed phrase as one whose distribution among documents is described by a Poisson distribution.[30] For such a phrase, the probability $P(k)$ that f_{jd}, the number of occurrences of phrase j in document d, is equal to k is given by

$$P(f_{jd} = k) = \frac{e^{-\lambda j}\lambda_j^{\,k}}{k!} \quad (9)$$

where the parameter λ_j is the average number of occurrences of j per document over the whole database. Therefore, we can find the probability p that the given phrase j occurs one or more times in d:

$$\begin{aligned} p &= 1 - P(f_{jd} = 0) \\ &= 1 - e^{-\lambda j} \end{aligned} \quad (10)$$

We denote by q (= $1 - p$) its complement, i.e., the probability that j does not occur in d. Let us consider an experiment that consists of N repeated independent Bernoulli trials with parameter p. Let $E(= N \cdot p)$ refer to the expected number of documents containing the phrase considered. If a phrase occurs multiple times in few documents, we say it has a tendency to clump. We measure the tendency to clump by how much the observed number of documents containing the phrase (i.e., *phrase frequency*) falls below the expectation E. For a given word pair we calculate the *P value*, i.e., the probability that *phrase frequency* would be less than or equal to that observed if it were generated by the Poisson distribution of equation (9).

$$P \text{ value} = \sum_{i=0}^{\textit{phrase frequency}} \binom{N}{i} p^i q^{N-i}. \quad (11)$$

Then the score s_6 is given by

$$s_6 = -\log(P \text{ value}) \quad (12)$$

Rationale: Intuitively, the occurrences of a term sensitive to content will have a greater tendency to clump than will those of a non-content-bearing term. This is common with names of things. Therefore, if the phrase considered carries content, we expect that the observed *phrase frequency* will be much less than E.

The scoring method s_6 is a measure of this clumping compared with a Poisson distributed phrase. For example, s_6 is 168.08 for the name "ulcerative colitis," which is highly specific, but s_6 is 0.65 for "common cancer," which is a general concept.

The same scoring methods discussed for the word pairs in the set M_2 can be applied to the word triples in the set M_3 with a slightly altered definition of co-occurrence. Given a word triple in the set M_3, phrase frequency is unchanged as the number of documents in MED96 that contain the word triple. However, for s_2, co-occurrence is the number of documents that contain both the first word and the second and third words contiguously as a word pair. The same definition of co-occurrence applies when computing s_5 and in equation (6), n_s is the number of documents that contain the second and third words as a word pair. The value of $phrase_{\text{end}}$ in the score s_3 in equation (3) is the number of phrases extracted from MED96 documents in which the given word triple in the set M_3 occurs at the right-hand end. The score s_4 likewise has the obvious interpretation where only $p(l_1l_2l_3)$ is changed to the distribution of letters $(l_1l_2l_3)$ appearing at the end of word triples in the set M_3.

Evaluation Method

Here we assume a given set of phrases M and a scoring method S that computes a real number for each phrase in the set M. The scoring method S allows us to rank the set M so that the phrases are in order of decreasing score. We also assume that we have available a golden set of good-quality phases G (this generally requires human judgment). The evaluation methods we use are measures of how well the scoring method S moves phrases in the set $M \cap G$ to the top of the listing of M by rank (the lowest ranks). In other words, we consider the phrases in $M \cap G$ to be the relevant phrases we are attempting to find in the set M. This allows us to view the problem as a retrieval problem and to apply some of the standard measures used in information retrieval science. In particular, we will apply recall and precision, which are the most commonly used measures in information retrieval.[31] Because recall and precision are generally defined for a given rank and the results are different for each rank considered, we will also use the 11-point average precision as a single summary measure for the complete ranking. We will further use interpolated recall-precision curves as a graphic way of viewing performance. The 11-point average precision and interpolated recall-precision curves are widely used in presenting the results of retrieval experiments.[32–34] Other measures are used in the information retrieval setting, such as the E-measure,[35] expected search length,[36] and relevance information.[37] While these measures have some advantages in specialized settings, they are less intuitive and less well known, and we feel they offer no advantage in our setting.

Let us assume that the number of phrases in M is N and that the phrases are represented by the list $\{ph_i\}_{i=1}^{N}$ indexed in rank order, where the order is that of decreasing score S. Further, let

$$value(ph_i) = \begin{cases} 1, \text{ if } ph_i \text{ is in the set } M \cap G \\ 0, \text{ otherwise} \end{cases} \quad (13)$$

Then the precision (P_r) and the recall (R_r) of S for the retrieval down to rank r is defined by

$$P_r = \frac{1}{r} \sum_{i=1}^{r} value(ph_i)$$

and

$$R_r = \frac{1}{\|M \cap G\|} \sum_{i=1}^{r} value(ph_i) \quad (14)$$

respectively. (Here $\|X\|$ denotes the number of elements in the set X.) In words, P_r is the fraction of phrases retrieved down to rank r that are in $M \cap G$, and R_r is the fraction of phrases in $M \cap G$ that are found in the retrieval down to rank r. Since the precision is usually high early in the ranks and becomes progressively lower at higher ranks, and since the recall is low at the early ranks but increases with increasing rank, it is possible to gain a useful picture of performance by graphing precision as a function of recall (a so-called recall-precision curve). However, precision does not always strictly decrease as one moves down the ranks. Because of this, it has become common to perform an interpolation on the precision value associated with a given recall level, in which that precision is replaced by any higher precision that may occur at a higher recall level. For example, if the precision 0.38 is calculated from equation (14) and the corresponding recall is 0.10, but a precision of 0.43 is found at a recall level of 0.20, then the value 0.38 is replaced by 0.43 as the accepted precision at recall level 0.10. In this way noise in the data may be reduced and the curve smoothed. We apply interpolation to obtain precision values at the 11 recall values of 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 percent. (Notice that the precision at 0 recall is the highest precision found at one or more ranks of retrieval, since at zero ranks P_r in equation (14) is undefined). These 11 recall-precision pairs are used to produce interpolated recall-precision curves. We also average the 11 precision values together to produce the 11-point average precision as an overall summary performance measure.

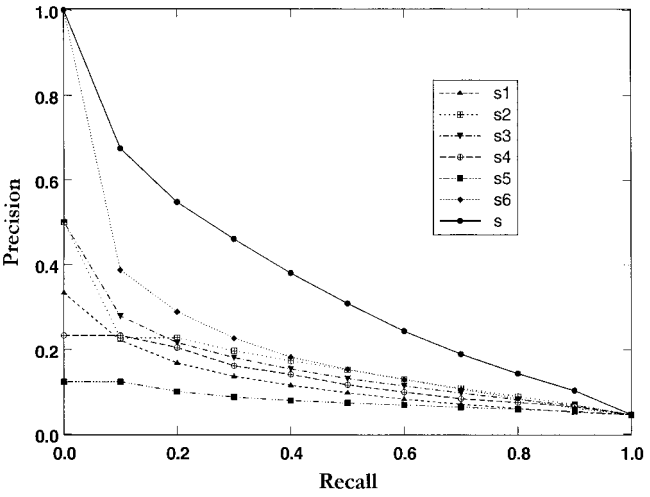

Figure 1 Recall-precision curves for the scoring methods applied to the set M_2. The 11 precision values are interpolated so that precision is a nonincreasing function of recall.

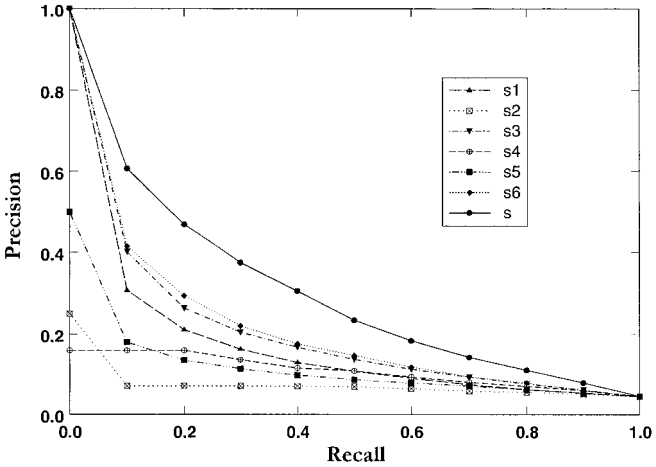

Figure 2 Recall-precision curves for the scoring methods applied to the set M_3. The 11 precision values are interpolated so that precision is a nonincreasing function of recall.

Results

Here we give the results of applying the scoring methods (defined under Scoring Methods) to the word pairs M_2 and word triples M_3 (described under Data Sources and Preparation). For the purpose of evaluation, the set of good phrases for M_2 is $G_2 = U_2 \cap M_2$ and the set of good phrases for M_3 is $G_3 = U_3 \cap M_3$. (The values U_2 and U_3 are defined under Data Sources and Preparation.) There are 26,131 phrases in G_2 and 9,234 phrases in G_3. The 11-point interpolated recall-precision curves for our scoring methods applied to the sets M_2 and M_3 are shown in Figure 1 and Figure 2. It can be clearly seen that each scoring method moves the relevant phrases toward the top of the lists. The 11-point average precision for each scoring method has been computed. Table 1 is the list of scoring methods and the 11-point average precision values on the set of word pairs M_2, and Table 2 provides corresponding data for the word triples in the set M_3. For either M_2 or M_3 it is possible to combine the different scoring methods to produce a composite score. For example, we may take the linear combination of the logarithm of each score s_i with a coefficient x_i and denote the resulting score as s:

$$s = \sum_{i=1}^{6} x_i \log(s_i). \quad (15)$$

Because the logarithm and the exponential functions are monotonically increasing functions of their arguments, s as defined by equation (15) is equivalent to its exponential for ranking purposes, and we would have obtained the same results if we have defined s as a product of the factors $s_i^{x_i}$. Ranking the phrases in order of a decreasing combined score s with the coefficients x_i where $i = 1 \ldots 6$, we can obtain the 11-point average precision. Through numeric study, we may seek the coefficients for which the combined score s gives the maximum 11-point average precision. We have iteratively maximized the 11-point average precision on one coefficient at a time. For example, x_1 is varied with the remaining five coefficients x_2, x_3, x_4, x_5, and x_6 fixed. This procedure is repeated until none of the coefficients can be altered to increase the 11-point average precision. We find that the combined score s with the coefficients $x_1 = -0.8$, $x_2 = 3.9$, $x_3 = 0.9$, $x_4 = 3.1$, $x_5 = 1.2$, and $x_6 = 1.7$—i.e.,

$$s = -0.8 \log(s_1) + 3.9 \log(s_2) + 0.9 \log(s_3) + 3.1 \log(s_4) + 1.2 \log(s_5) + 1.7 \log(s_6) \quad (16)$$

—gives the maximum 11-point average precision we are able to achieve for the word pairs in the set M_2. The result is listed in the final row of Table 1. Likewise, the combined score s with the coefficients $x_1 = -1.0$, $x_2 = 1.5$, $x_3 = 1.7$, $x_4 = 2.6$, $x_5 = 1.0$, and $x_6 = 2.2$—i.e.,

$$s = -1.0 \log(s_1) + 1.5 \log(s_2) + 1.7 \log(s_3) + 2.6 \log(s_4) + \log(s_5) + 2.2 \log(s_6) \quad (17)$$

—gives the maximum 11-point average precision, listed in the final row of Table 2, that is applicable to the word triples in the set M_3.

To get an idea how much each score contributes to the maximum 11-point average precision, we subtract its contribution from the combined score s. The resultant score (which is the optimal combined score s minus the contribution of an individual scoring method) and

Table 1 ■

Contribution of Individual Word-pair Scoring Methods to Combined Precision

Score	Precision s_i	Combined Score s less Individual Contribution	Precision $s - x_i \log(s_i)$
s_1	0.12	$s + 0.8 \log(s_1)$	0.36
s_2	0.17	$s - 3.9 \log(s_2)$	0.30
s_3	0.16	$s - 0.9 \log(s_3)$	0.34
s_4	0.13	$s - 3.1 \log(s_4)$	0.32
s_5	0.078	$s - 1.2 \log(s_5)$	0.33
s_6	0.24	$s - 1.7 \log(s_6)$	0.31
Combined s	0.37		

NOTE: Word-pair scoring methods are listed in column 1, and the corresponding precisions are given in column 2. Column 3 gives the optimal combined score with the contribution of the individual scoring method removed. The resultant drop in precision is shown by comparing the precision given in column 4 with the combined precision shown in the last row.

Table 2 ■

Contribution of Individual Word-triple Scoring Methods to Combined Precision

Score	Precision s_i	Combined Score s less Individual Contribution	Precision $s - x_i \log(s_i)$
s_1	0.20	$s + 1.0 \log(s_1)$	0.318
s_2	0.08	$s - 1.5 \log(s_2)$	0.31
s_3	0.17	$s - 1.7 \log(s_3)$	0.27
s_4	0.11	$s - 2.6 \log(s_4)$	0.28
s_5	0.13	$s - 1.0 \log(s_5)$	0.31
s_6	0.24	$s - 2.2 \log(s_6)$	0.29
Combined s	0.322		

NOTE: Word-triple scoring methods are listed in column 1, and the corresponding precisions are given in column 2. Column 3 gives the optimal combined score with the contribution of the individual scoring method removed. The resultant drop in precision is shown by comparing the precision given in column 4 with the combined precision shown in the last row.

the 11-point average precision value after applying it are listed in the third and the fourth columns of Table 1 for the word pairs in the set M_2. The third and the fourth columns of Table 2 provide corresponding data for the word triples in the set M_3. We can see that the precision of each scoring method does not directly add to the total precision found for the combined score s. For example, although the scoring method s_6 alone gives the largest 11-point average precision for both word pairs and triples, it does not make the largest contribution to the combined score s. This implies in particular that our scoring methods are not independent of each other.

In computing the combined scores given in equations (16) and (17), we have employed an optimization procedure to choose the coefficients. In such a computation there is always the possibility of overtraining, so that the results are applicable only to the particular data set on which we have done the training. We suspected that this would not be a problem in the current situation, because we are training only six parameters and are employing a very large number of data, namely, 304,057 MEDLINE documents. Furthermore, in actuality there are only five independent parameters, because ranking is not changed when one multiplies all scores by a constant. To test for overtraining, we randomly split the set of documents into disjoint data sets, MED^1 and MED^2, where MED96 = $MED^1 \cup MED^2$, $\|MED^1\| = 152{,}028$, and $\|MED^2\| = 152{,}029$. We then re-estimated the parameters for equation (16) on each subset independent of the other to produce the optimal 11-point average precision on that subset. The results are given in Table 3 along with the coefficients for the whole of MED96 for comparison.

Two things stand out here. First, we see that the coefficients obtained on MED^1 and MED^2 are almost identical. Second, we see that there is a significant difference between the coefficients obtained on the subsets and the whole database for scores s_1 and s_3. This suggests that database size might be a factor in these scores. To complete the comparison, we tested the effectiveness of each set of coefficients listed in Table 3 on the two subsets MED^1 and MED^2. The results are given in Table 4.

These results suggest that all three sets of coefficients have very close to the same effectiveness on MED^1 and MED^2. They are all within 1 percent of each other. This effectively rules out any significant overtraining. At the same time it suggests that the effectiveness of the composite scores is not very sensitive to the coefficients assigned to $\log(s_1)$ and $\log(s_3)$. While this is true, it is also true that we can degrade the composite quite drastically if we make these coefficients too large. A large coefficient for the contribution of a single score will cause this score to dominate the composite and the effectiveness of the composite to approach the effectiveness of the single score, as recorded in Tables 1 and 2.

Other Methods Tested

From our survey of the literature, the method of phrase identification that seems the closest in spirit to the method we have developed is that of Jones et al.[14] We tested their method on our task. The score of a phrase is given by a product $W \cdot F \cdot N^2$, where W is the sum of the frequencies of the individual words that make up the phrase, F is the phrase frequency, and N is the number of distinct non-stop words in the phrase. Since we consider only phrases without stop

words and since we apply the method to word doubles and word triples separately, the factor N^2 may be ignored for our purposes. With this scoring method, the 11-point average precision on M_2 is 0.06, and the 11-point average precision on M_3 is 0.107. Tables 1 and 2 show that in each case the result is not as good as s_1, which is equivalent to using F alone. This suggests that the factor W may be extraneous. As further support for this conclusion, we attempted to use the $W \cdot F \cdot N^2$ score to improve the combined scores for both word pairs and word triples; however, we were unable to improve our results in either case.

Part-of-speech tagging has been a popular method of extracting phrases from text for a variety of purposes.[25,38–43] We were naturally interested in including part-of-speech tagging as part of our system. To examine the possible benefits, we obtained access to the Xerox Part-of-Speech Tagger.[22] The tagger employed the SPECIALIST lexicon[44] and was trained on MEDLINE text. To apply tagging to our problem, we required a method of scoring word pairs and triples based on tagging. We give here the details for word pairs only. The trained tagger was used to tag the MED96 corpus. By this means each occurrence of a member of M_2 in MED96 has a tag pair associated from the tagging. We then constructed two lists of tag pairs. First, a set TM_2 of tuples (t_1t_2,m) was constructed, where t_1t_2 is any tag pair that occurs as a tag pair of an instance of some member of M_2 in MED96, and m is the number of different members of M_2 that occur at least once in MED96 with the tag pair t_1t_2. The set TG_2 is constructed on the basis of the same set of tag pairs as TM_2, the difference being that for $(t_1t_2,g) \in TG_2$, g is the number of different members of G_2 that occur at least once in MED96 with the tag pair t_1t_2. On the basis of these tag pair lists, we can estimate the important probabilities.

For any tag pair t_1t_2 with $(t_1t_2,m) \in TM_2$, we set

$$p(t_1t_2) = m/\|M_2\| \tag{18}$$

where $\|M_2\|$ denotes the size of the set M_2.

Likewise, for $(t_1t_2,g) \in TG_2$, we set

$$p(t_1t_2|good\ phrase) = g/\|G_2\| \tag{19}$$

Then, for any tag pair t_1t_2 seen in conjunction with a member of M_2, we associate an odds score

$$\begin{aligned} score(t_1t_2) &= p(good\ phrase|t_1t_2)/p(good\ phrase) \\ &= p(t_1t_2\|\ good\ phrase)/p(t_1t_2) \end{aligned} \tag{20}$$

Finally, if for any word pair $w_1w_2 \in M_2$ we let the set of tag pairs that correspond to all occurrences of w_1w_2 in MED96 be denoted by $T(w_1w_2)$, we define the score

Table 3 ■

Optimal Coefficients

Database	s_1	s_2	s_3	s_4	s_5	s_6
MED96	−0.8	3.9	0.9	3.1	1.2	1.7
MED[1]	−1.3	3.9	1.7	3.0	1.4	1.7
MED[2]	−1.3	3.9	1.8	2.9	1.4	1.5

Table 4 ■

Eleven-point Average Precision for Word Pairs

Source of Coefficients	Precision on MED[1]	Precision on MED[2]
MED96	0.411	0.409
MED[1]	0.412	0.412
MED[2]	0.412	0.413

$$score(w_1w_2) = \sum_{t_1t_2 \in T(w_1w_2)} score(t_1t_2)/\|T(w_1w_2)\|. \tag{21}$$

In an exactly analogous manner, a scoring function may be constructed for word triples in M_3. We might question the use of a straight average in equation (21). We tried a weighted average, in which the weight for $score(t_1t_2)$ was the number of occurrences of w_1w_2 in MED96 with tag pair t_1t_2. This actually gave worse results.

We tested the scoring functions for word pairs and word triples based on tagging in the same way the six scoring methods were evaluated before (see Evaluation Method). We found that the tagging score for word pairs produced an 11-point average precision of 0.166, while that for word triples produced an 11-point average precision of 0.167. These results are competitive with the results given for the scoring methods listed in Tables 1 and 2. An important question, however, is whether these methods add significantly to the methods already presented. We attempted to improve the composite scoring method for word pairs (described under Evaluation Method) by adding some fraction of log $score(w_1w_2)$ to it. We were able to improve the score only from 0.37 (bottom row Table 1) to 0.38. Likewise, we attempted to improve the composite score for word triples. Here we were marginally more successful, improving the composite score from 0.32 (bottom row Table 2) to 0.34.

Applications

We have applied the combined scoring methods derived as described under Evaluation Method to the MED96 data set described under Data Sources and Preparation. Of the 584,315 word pairs obtained from MED96 and ranked, the top one third were selected

as a set B_2 of sufficient quality to warrant consideration for inclusion in UMLS. Of the 206,522 word triples obtained and ranked, the top one half were selected as a set B_3 of sufficient quality to warrant consideration for inclusion in UMLS. The result is 157,867 two-word phrases in B_2 and 95,006 three-word phrases in B_3 that are not in UMLS.

While the methods described here can serve as a screen for the extraction of useful phrases, they can also form part of a system for marking useful phrases in text. Each marked phrase may then serve as a hyperlink to other texts that contain the same phrase. To illustrate the results of such processing, we have taken the sets B_2 and B_3, just defined, as the candidate phrases. The text of a document is broken into segments at punctuation marks and stop words. Each segment consisting of at least two words is examined, and the highest ranked member of B_2 and the highest ranked member of B_3 are selected for marking. If the two selected phrases overlap, only the word triple is marked, but if they do not overlap, both are marked. In some cases there is no word triple, and then the selected word pair is marked. In other cases no phrase scores high enough to be selected. Following is the result of such marking on a sample document from MED96.

> Title: **Impaired glucose tolerance** at **five-year follow**-up of **young men** with **borderline hypertension**.
>
> Abstract: **Recent studies suggest** that patients with **essential hypertension** have **impaired glucose tolerance** and are hyperinsulinemic compared with **normotensive subjects**. The aims of the study were (1) to follow **blood pressures** of 56 **young men** with **borderline hypertension** for 5 years, (2) to investigate **glucose tolerance** in these subjects, and (3) to determine the relation of insulin/**glucose metabolism** to structural vascular changes and hemodynamic patterns in **borderline hypertension**. METHODS: **Thirty-nine** young (age 22–34 years) **male subjects** with **borderline hypertension** (SBP 140–160 and or DBP 85–95 mmHg initially and 17 **normotensive control subjects** (SBP 110–130 and DBP 60–80 mmHg) participated in the study. **Blood pressure** was measured, a standard oral **glucose tolerance test** (OGTT) was performed, and glucose, insulin and **C-peptide** were determined before and 30, 60, 90 and 120 minutes after a standard 75-g **glucose load**. Post-ischemic forearm **vasodilatory responses** were examined by plethysmography. RESULTS: At follow-up, the **borderline hypertensives** had maintained significantly **higher blood pressures** than **control subjects**. **Borderline hypertensives** also had significantly **impaired glucose tolerance** compared to **control subjects**. The **insulin response** had a somewhat more sluggish descent, but did not **differ significantly** from the response of normotensives. The **C-peptide** response pattern resembled that of insulin, but **C-peptide** was **significantly elevated** after 120 min. On the whole group level, there were only weak relations of insulin to **blood pressure**. By contrast, **fasting insulin** and post-load **insulin levels** were **strongly correlated** with **body mass index**, the waist-**hip circumference ratio**, triglyceride, and both total and **LDL cholesterol**. Across the whole group, there were significant correlations between forearm **minimal vascular resistance** and **fasting insulin** ($r = +0.37$ $p = 0.007$) and insulin area-under-the-curve ($r = +0.28$ $p = 0.044$). However, Rmin was even more **strongly correlated** with **body mass index**, suggesting that this relationship was related to degree of obesity. CONCLUSION: **Borderline hypertension** in young men is a persistent condition which is associated with **impaired glucose tolerance** without hyperinsulinemia. This **finding suggests** that **impaired glucose tolerance** might be a more primary phenomenon in early hypertension devoid of lipid metabolic aberrations.

While the processing shown here is not perfect, it does mark most of the interesting and useful phrases that one might wish to follow as links to other documents. The main improvement that appears necessary is the elimination of throw-away phrases as **"five-year follow," "Recent studies suggest,"** and **"Thirty-nine."** Here, "five-year follow" should be "five-year follow-up," but "up" is on our stop list. As a rule, any phrase that contains the word "suggest" or "suggests" should be dropped. Likewise, phrases that are numbers are not useful as links. Thus, simple rules can be added to the system to improve the processing. Of particular note, syntactic parsing would appear capable of adding little to the analysis, since almost all the marked phrases appear syntactically reasonable.

Discussion

There are a number of limitations of the work reported here. One of these is our definition of a phrase. We exclude stop words from phrases. Because of this limitation, we cannot detect such phrases as "vitamin A," "hepatitis A," or "cancer of the lung." There are two things that can be done to help alleviate this problem. First, we can leave the letter "A" off the stop list while processing. This will generate significantly more phrases because the letter "A" is so prevalent in the language. If this is a burden on machine memory or disk space, we can even limit the processing to just those phrases that contain the letter "A" as one of the words in the phrase. As a second step, we can take good phrases that have been identified and use them to find additional useful phrases that contain stop words. For example, if we have identified the phrase

"lung cancer" as useful, we may then find "cancer of the lung" as a modified form. This could be accomplished in a general way by noting that the one is a rearrangement of the other with stop words added. It could also be accomplished, and with less error, by using a template that matches "UV" to "V of the U." We have also excluded phrases with more than three words. Relatively fewer useful phrases have four or more words in them. However, it would be useful to be able to identify phrases such as "high pressure liquid chromatography" and "left main coronary artery." We are currently examining methods by which we can extend our processing to obtain these longer phrases.

A second limitation of our approach is our dependence on the UMLS. First, we would like to point out that of the individual scoring methods presented under Scoring Methods, only s_4 depends for its derivation on the UMLS. The main use of the UMLS is to demonstrate that these scoring methods are effective in ranking useful phrases above non-useful phrases in a large list of phrases extracted from MED96. This is based on the assumption that the majority of phrases found in M_2 and M_3 are not useful and hence, if we can rank the phrases from the UMLS that occur in these lists near the top, we are succeeding in differentiating good phrases from bad. There still remains the question of the dependence of s_4 and the composite scores on the UMLS. Clearly there is a dependence, for we could not define these scores without the UMLS or some large set of good phrases. The question is whether the dependence is reasonable or not. We believe this question is answered in the affirmative by the results of the cross-validation testing presented under Results. We found the composite score (which includes a contribution from s_4) derived from one subset of MEDLINE, say MED^1, does not lose effectiveness when applied to a disjoint subset of MEDLINE, MED^2. Such results justify the application of the composite scores to new material in MEDLINE. We would, however, warn that the optimal choice of coefficients for MEDLINE might not be optimal for some other area of application.

Another aspect of our treatment that deserves comment is the fact that we derive the composite scores as a log linear sum of the individual scores. This is inspired by the inherent simplicity of the approach and by the wide success of log linear models in statistics. Although there is a strong similarity, our approach is somewhat different in that we seek to optimize the resultant ranking for retrieval rather than maximize the likelihood of the data. In other words, we seek to solve a slightly different problem. There is, at least theoretically, the possibility to do better modeling if we have detailed knowledge of the dependencies between the different individual scores. However, we are not in possession of this detailed knowledge, and this leads us to follow what is, for the present, a more feasible approach.

Finally, there is the question of whether some other methods of scoring may not prove useful for our task. To begin, we may ask why part-of-speech tagging does not prove more more successful in the task of identifying useful phrases. One issue is the performance of the Xerox tagger. In the initial description of the tagger, accuracy of more than 96 percent is claimed.[22] The implementation we used was trained on medical text, has been tested and used extensively in-house, and has performed well, and we see no reason to question the figure of 96 percent. If this figure is reasonably accurate, then we could not expect to see much improvement even if we were to use a tagger with 98 percent accuracy. We believe the lack of benefit we see from tagging stems from two sources. First, two of our scoring methods are based on properties of phrases that are syntactically important. The score s_3 is based on how often the phrase appears to have a potential head word as its last word, while score s_4 is based on the expectation that the last word in the phrase has a three-letter suffix that would be seen in a high-quality phrase. Since the Xerox tagger uses the suffix of an unknown word in predicting its ambiguity class, there is clearly overlap in the information used by s_4 and the tagger. While one could, in principle, test the level of dependence between s_4 and the tagger scoring, we have not attempted to do this.

The second point is that the task of identifying high-quality phrases is as much a problem of semantics as it is a problem of syntax. Tagging cannot help with the semantic problem, and thus the performance of tagging alone on the task at hand would seem to be limited. What we have said notwithstanding, there are other methods of part-of-speech tagging[20,23] that might yield different and more favorable results for our task. Other approaches that may prove useful are methods that require one to work from known good phrases to obtain related phrases that may also be of good quality. Examples include the work of Hersh et al.,[10] in which known good head words for phrases were used to locate numerous phrases built from them, and the work of Cooper and Miller[45] in locating good phrases that are lexical variations of MeSH terms (the lexical indexing system PostDoc) or that co-occur at a high level with MeSH terms (the statistical indexing system Pindex). While we do not question the effectiveness of these approaches, we have avoided them because they imply a strong correlation between what one has already given as good phrases and what one

can find with the methods. Our aim has been to accomplish a more general type of processing that would not bind us so strongly to prior knowledge.

In our current and ongoing work we are examining two ways of improving the system. First, as mentioned above, we would like to allow phrases longer than three words. We are seeking to do this by examining more closely the statistical dependency between words that occur in text, the idea being that a word that occurs at the left end of a phrase may belong there if the dependency is sufficiently strong. Second, we would like to find a way to score phrases more accurately as to how laden with content or subject matter they are. Bookstein et al.[46,47] have developed methods for this purpose that make use of the distribution of terms within a document. Those that are content bearing tend to be uneven in their distribution. Unfortunately, we have access only to titles and abstracts of documents and will have to take a different approach, more related to how the terms are distributed relative to other terms within the whole database.

The authors thank Alan Aronson and Jim Mork for making the trained Xerox tagger available for this study.

References ■

1. Funk ME, Reid CA, McGoogan LS. Indexing consistency in MEDLINE. Bull MLA. 1983;71(2):176–83.
2. Furnas GW, Landauer TK, Dumais ST, Gomez LM. Statistical semantics analysis of the potential performance of keyword information systems. Bell System Tech J. 1983;62: 1753–806.
3. Blair DC, Maron ME. An evaluation of retrieval effectiveness for a full-text document-retrieval system. Commun ACM. 1985;28(3):289–99.
4. Bates MJ. Subject access in online catalogs: a design model. J Am Soc Info Sci. 1986;37:357–76.
5. Bates MJ. Rethinking subject cataloging in the online environment. Libr Resources Tech Serv. 1989;33:400–12.
6. Gomez LM, Lochbaum CC, Landauer TK. All the right words: Finding what you want as a function of richness of indexing vocabulary. J Am Soc Info Sci. 1990;41:547–59.
7. Bates MJ. Indexing and access for digital libraries and the Internet: Human, database, and domain factors. J Am Soc Info Sci. 1998;49(13):1185–205.
8. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 1998;5(1): 1–11.
9. Joubert M, Fieschi M, Robert JJ, Volot F, Fieschi D. UMLS-based conceptual queries to biomedical information databases: an overview of the project ARIANE. J Am Med Inform Assoc. 1998;5(1):52–61.
10. Hersh WR, Campbell EH, Evans DA, Brownlow ND. Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. AMIA Annu Fall Symp. 1996:159–63.
11. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Molecular Biology of the Cell. 3rd ed. New York: Garland Publishing, 1994.
12. Luhn H. A new method of recording and searching information. Am Documentation. 1953;4:14–6.
13. Luhn HP. The automatic creation of literature abstracts. IBM J Res Develop. 1961;2:159–65.
14. Jones LP, Gassie EW Jr, Radhakrishnan S. INDEX. The statistical basis for an automatic conceptual phrase-indexing system. J Am Soc Info Sci. 1990;41(2):87–97.
15. Damereau J. An experiment in automatic indexing. Am Documentation. 1965;16(4):283–9.
16. Stone DB, Rubinoff M. Statistical generation of a technical vocabulary. Am Documentation. 1968;19:411–2.
17. Harter SP. A probabilistic approach to automatic keyword indexing. J Am Soc Info Sci. 1975;26:197–206.
18. Evans DA, Lefferts RG, Grefenstette G, Handerson SK, Hersh WR, Archbold AA. CLARIT TREC design, experiments, and results. In: Harman DK (ed). The First Text Retrieval Conference (TREC-1). Gaithersburg, Md: National Institute of Standards and Technology, 1992:251–86. NIST Special Publication 500-207.
19. Voutilainen A. NPtool: a detector of English noun phrases. Proceedings of the Workshop on Very Large Corpora; Columbus, Ohio; 1993.
20. Brill E. Automatic grammer induction and parsing free text: a transformation-based approach. Proceedings of the 31st Meeting of the Association of Computational Linguistics; Columbus, Ohio; 1993.
21. Brill E, Pop M. Unsupervised learning of disambiguation rules for part-of-speech tagging. In: Armstrong S, Church KW, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D (eds). Natural Language Processing Using Very Large Corpora. Dordrecht, The Netherlands: Kluwer Academic Press, 1999.
22. Cutting D, Kupiec J, Pedersen J, Sibun P. A practical part-of-speech tagger. Presented at the Third International ACL Conference on Applied Natural Language Processing; Trento, Italy; 1992.
23. Minsky, M, Haase K. CHOPPER. Cambridge, Mass: Machine Understanding Group, MIT Media Lab, 2000.
24. Fagan JL. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. J Am Soc Info Sci. 1989;40(2):115–32.
25. Lewis DD, Croft WB. Term clustering of syntactic phrases. In: Vidick J (ed). Proceedings of the 13th International Conference on Research and Development in Information Retrieval. Brussels, Belgium: Presses Universitaires de Bruxelles, 1990:385–404.
26. Lewis DD, Jones KS. Natural language processing for information retrieval. Commun ACM. 1996;39(1):92–101.
27. Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. Methods Inf. Med. 1993;32: 281–91.
28. Salton G, Wong A, Yu CT. Automatic indexing using term discrimination and term precision measurements. Info Proc Manage. 1976;12:43–51.
29. Allen J. Natural Language Understanding. Redwood City, Calif. Benjamin/Cummings, 1995.
30. Larson HJ. Introduction to Probability Theory and Statistical Inference. 3rd ed. New York: John Wiley, 1982.
31. Salton G. The state of retrieval system evaluation. Info Proc Manage. 1992;28(4):441–9.
32. Witten IH, Moffat A, Bell TC. Managing Gigabytes. 2nd ed. San Francisco, Calif: Morgan-Kaufmann, 1999.
33. Ponte JM, Croft WB. A language modeling approach to information retrieval. In: Croft WB, Moffat A, Rijsbergen CJ,

Wilkinson R, Zobel J (eds). SIGIR98. Melbourne, Australia: ACM Press, 1998:275–81.
34. Yang Y. An Evaluation of statistical approaches to text categorization. Info Retrieval. 1999;1(1):69–90.
35. van Rijsbergen CJ. Information Retrieval. 2nd ed. London, UK: Butterworths, 1979.
36. Cooper WS. Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. J Am Soc Info Sci. 1968;19:30–41.
37. Wilbur WJ. An information measure of retrieval performance. Info Syst. 1992;17(4):283–98.
38. Lewis D. An evaluation of phrasal and clustered representations on a text categorization task. In: Belkin N, Ingwersen P, Pejtersen AM (eds). Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark: ACM Press, 1992: 37–50.
39. Strzalkowski T, Vauthey B. Information retrieval using robust natural language processing. 30th Annual Meeting of the Association for Computational Linguistics (ACL); University of Delaware, Newark, Delaware; 1992. New Brunswick, NJ: ACL, 1992:104–11.
40. Strzalkowski T. Document indexing and retrieval using natural language processing. Proceedings of RIAO 94; Rockefeller University, New York. Paris, France: Jouve, 1994: 131–43.
41. Finch S, Partial orders for document representation: a new methodology for combining document features. In: Fox EA, Ingwersen P, Fidel R (eds). Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; Seattle, Washington; 1995. New York: ACM Press, 1995:264–72.
42. Jing Y, Croft WB. An association thesaurus for information retrieval. Proceedings of RIAO 94; Rockefeller University, New York. Paris, France: Jouve, 1994:146–60.
43. Bennett NA, He Q, Powell K, Schatz BR. Extracting noun phrases for all of MEDLINE. AMIA Annu Symp. 1999:671–5.
44. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc 18th Annu Symp Comput Appl Med Care. 1994:235–9.
45. Cooper GF, Miller RA. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. J Am Med Inform Assoc. 1998;5:62–75.
46. Bookstein, A., Klein ST, Raita T. Detecting content bearing words by serial clustering. In: Fox EA, Ingwersen P, Fidel R (eds). Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; Seattle, Washington; 1995. New York, ACM Press, 1995.
47. Bookstein A, Klein ST, Raita T. Clumping properties of content-bearing words. J Am Soc Info Sci. 1998;49(2):102–14.