Concept Extraction and Synonymy Management for Biomedical Information Retrieval

Colleen Crangle[1], Alex Zbyslaw[1], J. Michael Cherry[2], Eurie L. Hong[2]
[1]ConverSpeech LLC, Palo Alto, California
[2]Department of Genetics, Stanford University, California

Abstract

This paper reports on work done for the Genomics Track at TREC 2004 by ConverSpeech LLC in conjunction with scientists at the *Saccharomyces* Genome Database (SGD), the model organism database located at Stanford University, California. The rapidly increasing number of articles in the biomedical literature has created new urgency for software tools that find information relevant to specific information needs. We focused on two challenges in this work: the problems of synonymy (several terms having the same meaning) and polysemy (a term having more than one meaning), and the problem of constructing queries from information needs stated in natural language. We investigated the use of concept extraction for the second problem, relying on the limited statements of information need as the source of textual analysis. To minimize the problem of synonymy, we investigated the use of a language-oriented biomedical ontology and MeSH (Medical Subject Headings) for term expansion. Additionally, to minimize the problem of polysemy, we used extracted concepts to analyze and rank the documents returned by a search. We submitted two sets of results to TREC for evaluation, the first one produced automatically, the second derived from the first by making specific kinds of changes in the query and ranking methods. The mean average precision (MAP) for the automatic result was lower than the median of the 37 submitted runs overall; however, desirable results were obtained for mean average precision at 10 and 100 documents for almost half the topics. The MAP for the derived result was higher than the median, a desirable result.

Background

NEED. The rapidly increasing number of articles in the biomedical literature has created new urgency for software tools that find information relevant to specific information needs. The Text Retrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, supports large-scale evaluation of text retrieval methodologies. The TREC 2004 Genomics track contained a task, called the Ad hoc information retrieval task, that consisted of 50 specific information needs collected from interviews with biomedical scientists. Documents relevant to these needs had to be located within a 10-year subset of the MEDLINE bibliographic database and the results sorted according to their estimated relevance. This paper reports on work done for the Ad hoc task by ConverSpeech LLC in conjunction with scientists at SGD, a scientific database of the molecular biology and genetics of the yeast *Saccharomyces cerevisiae* located at Stanford University, California [1].

PROBLEMS. *The language of biomedical text.* The language of biomedical texts, like all natural language, is complex in structure and morphology (the basic units of meaning) and poses problems of synonymy (several terms having the same meaning), polysemy (a term having more than one meaning), hypernymy (one term being more general than another), and hyponymy (one term being more specific than another), among others. We investigated the use of our language-oriented biomedical ontology, called BioMedPlus, along with MeSH (Medical Subject Headings) terms to provide synonyms to produce an expanded query. In doing so, the problem of polysemy was exacerbated, which we then also addressed.

Natural-language information needs. The second major challenge we addressed was how to transform the natural-language descriptions given by the interviewed biomedical scientists into queries that can be submitted to an information retrieval system. Concept extraction is the process of deriving terms from natural-language text that are considered representative of what the text is about. The terms are natural-language words and phrases, which may or may not themselves occur with the original text. We investigated the use of concept extraction to produce both a set of terms for the initial query and a set of terms for analyzing the documents returned by the initial query, ranking them according to their estimated relevance. This ranking helps minimize the problem of polysemy.

Limited sources for textual analysis. Although inverse document frequency methods have long been used to assess the retrieval quality of a term [15], we wanted to evaluate a way of finding good terms when a document set may not be available, or may be changing too rapidly, for the tf*idf method, or when all that is available is a short piece of text. This interest led to our use of concept extraction for this information retrieval task.

HYPOTHESIS. The primary evaluation measure for the task was mean average precision (MAP) defined as the average precision at each point a relevant document is retrieved. Recall and precision are calculated as usual, with recall being the percentage of relevant documents returned by the system at the point of measurement, and precision being the percentage of the documents returned by the system at the point of measurement that are actually relevant. The hypothesis of our study was that our concept extraction methods and our approach to synonymy and polysemy would yield desirable MAP results.

Results

We submitted two sets of results to TREC for evaluation, one labeled ConversAuto and the other ConversManu. The first was produced automatically, with no human intervention. The second was derived from the first after consultation with our collaborators at SGD. We added an additional automatically constructed query term for all topics and made three additional kinds of changes in the query and analysis (ranking) terms that proved successful. The mean average precision (MAP) for ConversAuto was 0.2013 and for ConversManu 0.2931. The MAP for ConversManu was higher than the median 0.241 of all the 37 submitted runs, a desirable result. The MAP for ConversManu was lower, but for almost half the topics the mean average precision at 10 and 100 documents surpassed or equaled the corresponding medians.

Materials and Methods

MEDLINE. A typical MEDLINE record consists of fields for author name(s), title, abstract, date of publication (DP field), MeSH terms (MH field with Medical Subject Heading terms selected by bibliographic and subject specialists), chemical terms (RN field with terms selected by bibliographic and subject specialists), date the MEDLINE record was completed (DCOM field), and so on. The 10-year window used to select MEDLINE records was from 1994 to 2003 inclusive, as identified by DCOM. Within this subset, therefore, there were some records for articles published before 1994; the DCOM field notes the date the article's citation was completed for MEDLINE, not the article's date of publication. The subset, in fact, had 2,814 (0.06%) articles published prior to 1980, 8,388 (0.18%) articles published prior to 1990, and 138,384 (3.01%) articles published prior to 1994. The remaining 4,452,624 (96.99%) articles were published within the 10-year period of 1994-2003. Of the total 4,602,210 articles, approximately 75% had abstracts; the remaining 25% lacked abstracts.

TOPICS. The topics consisted of the following three fields (plus an identification field used for processing):

- Title, an abbreviated statement of the information need
- Information need, a full statement the information need
- Context, background information to place the information need in context.

In addition to the 50 topics used for the runs submitted to TREC, an additional five sample topics were made available for early experimentation. One is displayed in Figure 1.

Title	pBR322 used as a gene vector
Need	Find information about base sequences and restriction maps in plasmids that are used as gene vectors
Context	The researcher would like to manipulate the plasmid by removing a particular gene and needs the original base sequence or restriction map information of the plasmid

Figure 1 - Sample topic

BIOMEDPLUS ONTOLOGY. The ConverSpeech ontology, BioMedPlus, is a federated, language-oriented ontology constructed from LocusLink [4], GO [5], KEGG [6, 7], and SGD [1], and modeled on WordNet [2,3], a widely used general ontology for the English language. It includes a vocabulary (which promotes a standard way of naming the concepts of the domain) and a system of hierarchical and other relations between and among the concepts and the vocabulary items. It also includes definitions of the vocabulary items. Concepts in BioMedPlus are represented as synonym sets, which are sets of words or phrases that express the same meaning or refer to the same biomedical entity in at least one context. The vocabulary in BioMedPlus is therefore divided into sets of synonyms, each representing a single underlying concept. Information on synonyms is readily available in the biomedical sources from which the ontology is built. Each synonym set may have one or more relationships to other synonym sets. For example, a hypernymy relationship indicates that one concept is a kind of (or subordinate to) the other concept. Glucose metabolism, for example, is a kind of hexose metabolism, which in turn is a kind of monosaccharide metabolism. BioMedPlus can also store non-hierarchical relationships, such as GO associations. The same word or phrase may occur in more than one source ontology; however each usage of the word or phrase may have a different meaning. To differentiate between these meanings, separate uses of a word or phrase are assigned separate sense number, which are simple integers. For example, the first use of "SFD" is assigned the sense number 1 (written as "SFD#1"), the second usage assigned sense number 2 ("SFD#2"), and so on.

CONCEPT EXTRACTION. Concept extraction is the process of deriving terms from natural-language text that are considered representative of what the text is about. The terms are natural-language words and phrases, which may or may not themselves appear in the original text. We began by counting as concepts those words and phrases in the Title, Need, and Context that were:

- also entries in the ConverSpeech BioMedPlus ontology,
- collocations found through statistical analysis of some combination of Title, Need, and Context, using the method of likelihood ratios [15], and
- words and phrases that were neither part of general English (as determined by our adapted WordNet model) nor entries in the BioMedPlus ontology. This rule aimed to capture those biomedical terms that have not yet found their way into any of the sources for BioMedPlus but are not regular English words.

Words and phrases that were simple pluralizations of words and phrases already extracted were not considered separate concepts. Words and phrases on a Stop List, currently compiled heuristically, containing terms judged too general to be of interest (e.g., "gene") were eliminated. No synonyms were used at this stage; the concepts extracted so far were always themselves words or phrases appearing in the Title, Need, or Context. Trial and error experimentation using the five sample topics led us to use the following concepts for the initial query.

Let C_1 be the set of concepts extracted from the Title, Need, and Context.

Let C_2 be the set of concepts extracted from the Title and Need.

Then let Q_1, Q_2, and Q_3 be sets that contain terms for the initial query, where

Q_1 contains every term from C_2 that is from GO or KEGG; that is, the Title and the Need concepts that correspond to GO or KEGG terms,

Q_2 contains every remaining term from C_2 that appears in the Title but is not part of general English (as determined by our adapted WordNet), and

Q_3 contains every remaining term in C_2.

For analyzing and ranking the returned documents according to their estimated relevance, we defined an additional set A_1 as follows:

A_1 contains every term from C_1 which did not appear in Q_1, Q_2 or Q_3, that is, the concepts provided by the Context that were not already in the other concept sets.

These sets were defined to give greater weight to the Title and Need and to words and phrases derived from GO and KEGG. The initial query was then constructed as follows:

(Boolean OR of all terms in Q1) AND (Boolean OR of all terms in Q2) AND
(Boolean OR of all terms in Q3)

Figure 2 shows the concepts, initial query and ranking concepts for the sample topic in Figure 1.

C_1	C_2
base sequence gene vector pBR322 plasmid restriction map	gene vector pBR322 plasmid

Q_1: <empty>
Q_2: pBR322
Q_3: plasmid, gene vector

Initial query: (pBR322) AND (plasmid OR gene vector)
Ranking concepts: base sequence, restriction map

Figure 2 - Concepts, initial query, and ranking concepts for sample topic

For the derived queries used to produce ConversManu, we constructed an additional query term, Q_{11}, for each topic. It consisted of every remaining term from C_2 that is a species name. The derived query was then built by adding the terms in Q_{11} as follows:

(Boolean OR of all terms in Q_1) AND (Boolean OR of all terms in Q_2) AND
(Boolean OR of all terms in Q_3)
AND (Boolean OR of all terms in Q_{11}).

There are many different ways of using the Title, Need, and Context concepts in constructing the query, and many different logical forms the query could take. Without further analysis to see what works best, and possibly why, we make no claim about the specific form of the queries. Our main interest was in seeing if the concepts extracted appeared to pick out significant terms for searching and ranking that gave reasonable results.

SEARCH ENGINE. The query was processed by a proprietary search engine that uses the PubMed e-utilities [8]. First, however, each term was run through the BioMedPlus ontology to find synonyms. An expanded search query was built with the synonyms added in using the Boolean OR where appropriate. When this query was passed to PubMed via the e-utilities, a further step of query transformation took place. MeSH translations were provided wherever one existed [9] and, again, using the Boolean OR, a final query was constructed for searching MEDLINE. The MeSH translations often appended the search delimiter [MH] to the MeSH term, ensuring that the term would be matched only against the same term in the MH field of the MEDLINE record. At this stage, the concepts extracted and constructed from the Title, Need, and Context are represented by many terms that do not appear in the original natural-language description. Additional terms have come from synonyms and MeSH terms. Figure 3 shows one synonym that was added for topic 17 and two MeSH translations that were provided for topic 15.

Term	Synonym or MeSH translation
Trp53	transformation related protein 53
atpase	adenosinetriphosphatase [MH]
binding	pharmacokinetics [MH]

Figure 3 – Synonyms or MeSH translations

Because PubMed does not allow searching by DCOM—the date field used to create the base document set—our queries were limited instead by the EDAT field, the date the citation was added to the PubMed database, and the results were pruned against the list of PubMed IDs known to be in the base document set.

Typically, adding in biomedical synonyms introduces a great many terms, many of which in different contexts are not synonyms with the original term at all. For example, the following are all synonyms for "CGI-11":

ATP6V1H, ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H,
SFD, SFDalpha, SFDbeta,VMA13.

But the term "SFD," an abbreviation for "sub fifty-eight-kDa doublet" or "sub-fifty-eight-kDa dimer," is also widely used as an acronym in biomedical texts and has several dozen other meanings as a result. Adding this term to the search will pull in many irrelevant articles. By matching the concepts extracted from the original natural-language statements of need to those in the returned documents, we can score the documents based on their fit with the original concepts and so rank them by estimated relevance. Documents that have nothing to do with the original statement of need typically score so low that they essential fall out of contention.

ANALYZING THE DOCUMENTS FOR RANKING. Each document found by PubMed was analyzed against both the query terms (Q_1, Q_2, Q_3) and the analysis terms (A_1) and given a score. This analysis was performed on a local version of the PubMed documents as supplied to TREC participants. The terms were matched using a NORM-like algorithm adapted from the UMLS NORM procedure [10]. The algorithm works as follows. Given a phrase that we are trying to match:

- The words in that phrase are converted to lower-case.
- Words from the Stop List are discarded.
- The words are stored, for matching in any order.
- The words are also passed through the Porter Stemmer, a widely available algorithm for stemming English words so that words with the same stem can be treated identically. E.g., "connector" and "connections" both stem to "connect."

Biomedical terms frequently appear in variant forms. We took lexical variants into account by using the following rules to conflate variants and have them count as the same term:

- Hyphenated expressions such as "IL-12" are treated as two words ensuring that we can match such expressions when an author and the ontology disagree about the hyphenation. E.g., "IL 12" in the text will match "IL-12" in the ontology and vice versa.
- Any expression of the form {numbers}{letters} or {letters}{numbers} is treated as two separate words. E.g., "TH1" is considered two words, "TH" and "1", and "57kDa" is two words "57" and "kDa."

- All other internal punctuation is treated as a word break. E.g., “DUR1,2” would be two words “DUR1” and “2”. However, because these rules are cumulative, the previous rule also applies to “DUR1” which is treated as two words “DUR” and “1.”

For scoring the documents returned by the search, we made the following matches using the title, abstract, RN (chemical terms) field and MH (MeSH terms) field of the MEDLINE record. We scored each match, giving the greatest weight to matches in the title and the last two sentences of the abstract.

1. For each term in the query, including any synonyms, full matches in the title and in the last two sentences of the abstract scored 16 for the first occurrence and 8 for each subsequent occurrence. Full matches elsewhere (i.e., rest of abstract, MH, RN) scored 8 for the first occurrence and 4, 2, and 1 for the second, third and subsequent matches.
2. For each analysis term extracted from the topic, full matches in the title and in the last two sentences of the abstract scored 20 for the first occurrence and 10, 5, 3, and 1 for the second, third, fourth, fifth and subsequent matches. Full matches elsewhere (i.e., rest of abstract, MH, RN) scored 10 for the first occurrence and 5, 3, and 1 for the second, third and subsequent occurrences.

The title and last two sentences of the abstract are scored first, then the rest of the abstract and the MH and RN fields. So, for example, a term occurring in the title, last sentence of the abstract, and in the body of the abstract has its first and second occurrences in the highest scoring parts of the document. Occurrence is counted anywhere. So if a query term "t bet" matches in the title and in the middle of the abstract it would score 16 (first occurrence in title) + 4 (second occurrence elsewhere). This scoring was determined through trial and error and we make no particular claim for its general applicability. It merely served as some reasonable way of assessing the value of the extracted concepts for analyzing and ranking the returned documents.

RELEVANCE JUDGMENTS. Relevance judgments for all runs submitted to TREC were done using the conventional “pooling method.” That is, the top-ranking documents from each official run were pooled and given to an individual (blinded to the query statement and participant from whom they came) who judged relevance. The pools were built by collecting one run from each of the 27 participating groups, taking the top 75 documents for each topic and eliminating the duplicates to create a single pool for each topic. The average pool size (average number of documents judged per topic) was 976, with a range of 476-1450. Given that neither the pools nor the documents judged relevant are necessarily true subsets of the relevant documents, we produced Figure 4 to help us evaluate our results. The interpretation of the sets in Figure 4 is given in Table 1.

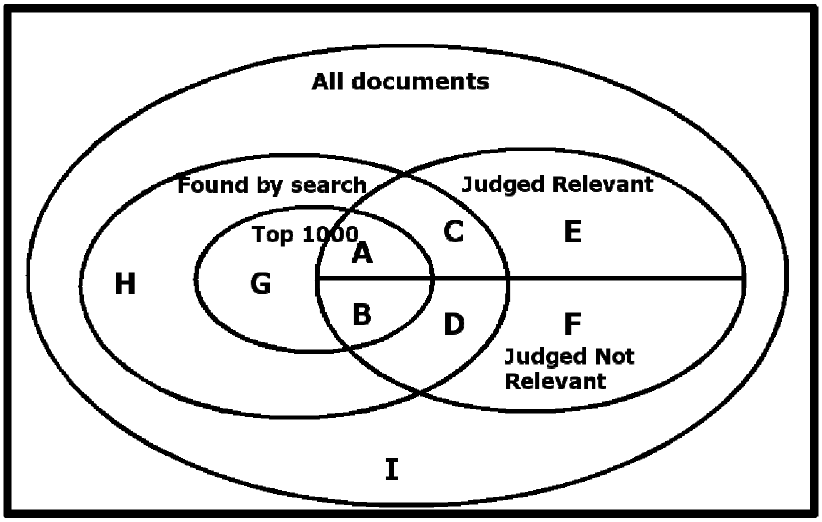

Figure 4 - Theoretical results possible for each topic

The set labeled "Found by search" contains all the documents returned by our search for that topic. Up to 1000 documents could be submitted for each topic in a run. The set labeled "Top 1000" represents the subset of found documents that were submitted. If fewer than 1000 documents were found, then all those found were submitted, and the set "Top 1000" actually contained fewer than 1000 and was identical to the set "Found by search." In such a case, sets C, D, and H will be empty. The set formed by the union of "Judged Relevant" and "Judged Not Relevant" is the pool that was submitted to the judge.

Set	Label	Found	Submitted	Relevant	Comment
A	Positive	Yes	Yes	Yes	What we are after! Contains relevant documents that we found.
B	False Positive	Yes	Yes	No	Contains documents we would like to score lower or not be found.
C	False Negative A	Yes	No	Yes	Contains documents we would like to score higher.
D	Negative A	Yes	No	No	Contains documents we found that were correctly eliminated by scoring.
E	False Negative B	No	No	Yes	Contains documents that we missed through an overly restrictive search.
F	Negative B	No	No	No	Contains documents we did not find that were irrelevant but some other run found them relevant.
G	False Positive Or Positive	Yes	Yes	Unknown	Contains documents we found that did not make it into the pool. They might be relevant documents missed or ranked too low by others, or (more likely) were not relevant but we found them and ranked them too high when they were ranked very low or passed over by others.

Set	Label	Found	Submitted	Relevant	Comment
H	False Negative Or Negative	Yes	No	Unknown	Contains documents we found that did not make it into the pool, and didn't make it into our submission wither. They might be relevant documents missed or ranked too low by others, or (more likely) were not relevant and we found them but correctly ranked them low when they were ranked very low or passed over by others.
I	All documents in search set	No	No	Unknown	Everything.

Table 1 - Explanation of theoretical results possible for each individual topic

Discussion

The results of greatest interest to us were for those topics that saw a significant improvement from the automatic run to the manual or derived run. Examining the changes made, we identified four that were particularly successful: (1) automatically adding to the query an additional conjunctive term for the species name, as discussed earlier; (2) adding query terms to the analysis terms, or visa versa; and (3) adding a query term that was needed because of the inadequacy of the stemming algorithm. For several topics, the terms "apoptosis" and "apoptotic," for example, did not stem to the same form. Adding "apoptotic" gave better results and better rankings. We are now investigating adapting or building a stemmer that is specifically attuned to biomedical language.

One other general change that was of interest was (4) that for some topics the automated method produced no query or analysis terms at all. This problem was responsible most often for poor results in the automated run. It resulted from an extra step in the concept extraction process that was not laid out in the earlier discussion. That is, in this process all the possible concepts are pooled along with their immediate hypernyms and definitions (if they exist) and formed into a network that relates individual words to phrases that they appear in. The network also tracks the frequency of occurrence of each word or phrase. Only those concepts that, through their degree of relatedness within the network, reach a predefined significance threshold are counted as final concepts. For example, for topic 18, before the final assessment of the terms' place in its network of related concepts, the following emerged as likely concepts: cell, cell cycle, Gis4, metabolism, and yeast. Afterwards, no concepts were considered significant. The query term proposed by the biologist was "Gis4" and the analysis terms "cell cycle," "metabolism," and "yeast carbon pathways." The terms that proved useful for analysis were, in fact, "Gis4" and "metabolism." So it turns out the concept extraction method was producing good concepts; they were often dropped, however, on the incorrect judgment that they did not reach threshold significance.

Title	Gis4
Need	Properties of Gis4 with respect to cell cycle and/or metabolism.
Context	It is possible that Gis4 plays a role between cell cycle and yeast carbon pathways and that there is a link between cell cycle and metabolism. A relevant document is one that supports or refutes this hypothesis with regard to the properties of Gis4 in one or both processes.

Figure 5 – Topic 18

Although this final step in concept extraction has proved useful for other tasks, not in the biomedical domain [11], we would not use it again, but would revert to the more straightforward selection of concepts that has been used for other biomedical tasks [12, 13].

References

[1] Dolinski, K., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R., Oughtred, R., Theesfeld, C. L., Binkley, G., Lane, C., Schroeder, M., Sethuraman, A., Dong, S., Weng, S., Miyasato, S., Andrada, R., Botstein, D., and Cherry, J. M. "Saccharomyces Genome Database" ftp://ftp.yeastgenome.org/yeast/ (July 2004).

[2] Christiane Fellbaum (Ed.) WordNet: An Electronic Lexical Database. 1998. Bradford Books.

[3] Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. Introduction to WordNet: an on-line lexical database.' International Journal of Lexicography 3 (4), 1990, pp. 235 - 244.
ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps
http://www.cogsci.princeton.edu/~wn/

[4] National Center for Biotechnology Information. http://www.ncbi.nih.gov/locuslink/, 2004.

[5] Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium (2000) Nature Genet. 25: 25-29.

[6] Kanehisa, M.; A database for post-genome analysis. Trends Genet. 13, 375-376 (1997). [pubmed]

[7] Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000). [pubmed]

[8] See http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html.

[9] National Library of Medicine (MeSH home page) http://www.nlm.nih.gov/mesh/, 2004.

[10] Humphreys, BL and DA Lindberg and HM Schoolman and GO Barnett (1999). The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc.

[11] Crangle, C.E., "Text summarization in data mining." In D. Bustard, W. Liu, and R. Sterritt (Eds.): Soft-Ware 2002, LNCS 2311, pp. 332-347, 2002. Springer-Verlag

Berlin Heidelberg. Proceedings of Computing in an Imperfect World, First International Conference, Belfast, Northern Ireland, April 2002. (pp. 332-347).

[12] Crangle CE. Using the Gene Ontology for text data mining: a case study with human disease genes in yeast. GO Users Meeting, Stanford University, January 2004. http://www.geneontology.org/meeting/Stanford_GO_Program2004.

[13] Crangle, C, Sopchak, L. An ontology improves text information access: A case study using human disease genes in yeast. Biomedical Information Science and Technology Initiative (BISTI) 2003 Symposium. Digital Biology: The Emerging Paradigm. November 6-7, 2003. NIH, Bethesda, Maryland. http://www.bisti.nih.gov/2003meeting/abstracts/

[14] C.D. Manning and H. Schutze. Foundations of Statistical Natural Language Processing. MIT press, Cambridge, MA, 1999.

[15] Ricardo A. Baeza-Yates , R. Baeza-Yates , Berthier Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1999.