SYNTACTIC APPROACHES TO AUTOMATIC BOOK INDEXING

Gerard Salton
Department of Computer Science
Cornell University
Ithaca, NY 14853

ABSTRACT

Automatic book indexing systems are based on the generation of phrase structures capable of reflecting text content. Some approaches are given for the automatic construction of back-of-book indexes using a syntactic analysis of the available texts, followed by the identification of nominal constructions, the assignment of importance weights to the term phrases, and the choice of phrases as indexing units.

INTRODUCTION

Book indexing is of wide practical interest to authors, publishers, and readers of printed materials. For present purposes, a standard entry in a book index may be assumed to be a nominal construction listed in normal phrase order, or appearing in some permuted form with the principal term as phrase head. Cross-references ("see" or "see also" entries) between index entries are also normally used in the index. Excerpts from two typical book indexes appear in Fig. 1.

Attempts have been made over the years to mechanize the book indexing task, based in part on the occurrence characteristics of certain content words in the document texts [Borko, 1970], and in part on more ambitious syntactic methodologies. [Dillon, 1983] However, as of now, completely viable automatic book indexing methods are not available. Two main research advances may, however, lead to the development of improved automatic book indexing procedures. These include the generation of advanced syntactic analysis procedures, capable of analyzing unrestricted English texts, as well as the construction of powerful automatic indexing systems using sophisticated term weighting systems to assess the importance of the indexing units. [Salton 1975a, 1975b] By joining the available linguistic procedures with the available know-how in automatic indexing, satisfactory book indexing systems may be developed.

AUTOMATIC PHRASE CONSTRUCTION

Book indexing systems differ from standard automatic text indexing systems because complex, multi-word phrases are normally used for indexing purposes rather than the single term entries that are preferred in conventional automatic indexing systems. The phrase generation system described in this note is based on an automatic syntactic analysis of the available texts followed by a noun-phrase identification process using parse trees as input and producing lists of nominal constructions. The parsing system used in this study is based on an augmented phrase structure grammar, and was originally designed for use in the EPISTLE text-critiquing system.[1] (Heidorn, 1982, Jensen, 1983)

A typical document abstract is shown

This study was supported in part by a grant from OCLC Inc., and in part by the National Science Foundation under grant IRI-87-02735.

[1] The writer is indebted to the IBM Corporation and to Dr. George Heidorn for making available the PLNLP parsing system for use at Cornell University.

in Fig. 2, and the output produced by the syntactic analysis program for sentence 2 of the document is shown in Fig. 3. It may be noted that the syntactic output appears in the form of a standard phrase marker, the various levels of the syntax tree being listed in a column format from left to right. During the analysis, a head is identified for each syntactic constituent, identified by an asterisk (*) in the output. Thus in Fig. 3, the VERB is the main head of the sentence; the head of the noun phrase preceding the main verb is the NOUN representing the term "operations", etc.

The phrase formation system used in this study builds two-term phrases by combining the head of a constituent with the head of each constituent that modifies it. (Fagan 1987a, 1987b) For the sample sentence of Fig. 3, such a strategy produces the phrases

development	-	exception
dictionary	-	development
negative	-	dictionary
system	-	operations

In the phrase output, the dependent term is listed first in each case, followed by the governing term. Note that the phrase generation system identifies apparently reasonable constructions such as "dictionary development" and "system operations", but not the unwanted phrases "exception operations" or "exception systems".

AUTOMATIC PHRASE ASSIGNMENT

An automatic phrase construction system generates a large number of phrases for a given text item. Fig. 4 lists all the phrases produced for the abstract of Fig. 2. Phrases occurring in the document title are identified by the letter T, and phrases obtained more than once for a given document are identified by a frequency marker (2) in Fig. 4. The output of Fig. 4 could be used directly in a semi-automatic indexing environment by letting the user choose appropriate index entries from the available list. The standard entries from the figure might then be manually chosen for indexing purposes by the document author, or by a trained indexer.

In a fully automatic indexing system, additional criteria must be used, leading to the choice of some of the proposed phrase constructions, and the rejection of some others. The following criteria, among others, may be useful:

- For sentences that produce more than one acceptable syntactic analysis output, all analyses except the first one may be eliminated; (in the Heidorn-Jensen analyzer multiple analyses are arranged in decreasing order of presumed correctness).
- Phrases consisting of identical juxtaposed words ("computations-computation" in Fig. 4) may be eliminated.
- Phrases consisting of more than two words (e.g. "document-retrieval-system") may be given preference in the phrase assignment process.
- Phrases occurring in document titles, and/or section headings may be given preference.
- Noun-noun constructions might be given preference over adjective-noun construction.

A further choice of phrases, as well as a phrase ordering system in decreasing order of apparent desirability, can be implemented by assigning a *phrase weight* to each phrase and listing the phrases in decreasing weight order. Two different frequency criteria are important in phrase weighting:

- The frequency of occurrence of a construct in a given document, or document section, known as the term frequency (tf)
- The number of documents, or document sections, in which a given construct occurs, known as the document frequency (df).[2]

[2] For book indexing purposes, a book can be broken down into sections, or paragraphs; the term frequency and document frequency factors are then computed for the individual book components.

The best constructs for indexing purposes are those exhibiting a high term frequency, and a relatively low overall document frequency. Such constructs will distinguish the documents, or document sections, to which they are assigned from the remainder of the collection. The corresponding term weighting system, known as tf·idf is computed by multiplying the term frequency factor by an inverse document frequency factor.

Fig. 5 shows selected phrase output based in part on the use of automatically derived term weights. The top part of the figure contains the automatically derived constructs containing more than two terms. These might be used for indexing purposes regardless of term weight. In addition, the two-term phrases whose term frequency exceeds 1 in the document might also be used for indexing purposes. This would add the 9 phrases listed in the center portion of Fig. 5.

Some of the phrases with tf > 1 have either a very high document frequency (125 for "retrieval system") or a very low document frequency of 1, meaning that the phrase occurs only in the single document 659. In practice, a reasonable indexing policy consists in choosing phrases for which tf $> k_1$ and $k_2 <$ df $< k_3$ for suitable parameters k_1,k_2, and k_3. When these parameters are set equal to 1, 1 and 100, respectively, the 5 phrases identified by asterisks in Fig. 5 are chosen as indexing units.

The bottom part of Fig. 5 shows a ranked phrase list in decreasing order according to a composite (tf × idf) phrase weight. Using such an ordered list, a typical indexing policy consists in choosing the top n entries from the list, or choosing entries whose weight exceeds a given threshold T. When T is chosen as 0.1, the 12 phrases listed at the bottom of Fig. 5 are produced. It may be noted that most of the terms listed in Fig. 5 appear to be reasonable indexing units.

In a practical book indexing system, a phrase classification system capable of determining relationships between similar, or identical, phrases becomes useful. Such a phrase classification then leads to the choice of canonical representations for each group of equivalent phrases, and to the assignment of "see" and "see also" references. Phrase relationships can be determined by using synonym dictionaries and various kinds of phrase lists. In addition, attempts have also been made to use the term definitions contained in machine-readable dictionaries to construct hierarchies of word meanings. (Walker, 1987; Kucera, 1985; Chodorow, 1985) The automatic construction of phrase classification systems remains to be pursued in future work.

REFERENCES

Borko, H., 1970, Experiments in Book Indexing by Computer, *Information Storage and Retrieval*, 6:1, 5-16.

Chodorow, M.W., Byrd, R.J., and Heidorn, G.E., 1985, Extracting Semantic Hierarchies from a Large On-Line Dictionary, *Proceedings of 23rd Annual Meeting of the Associations for Computational Linguistics*, Chicago, IL.

Dillon, M. and McDonald, L.K. 1983, Fully Automatic Book Indexing, *Journal of Documentation*, 39:3, 135-154.

Fagan, J.L., 1987a, Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods, Doctoral Dissertation, Cornell University, Technical Report 87-868, Department of Computer Science, Cornell University, Ithaca, NY.

Fagan, J.L., 1987b, Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-Syntactic Methods, *Tenth Annual ACM/SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA, ACM, NY, 1987.

Heidorn, G.E., Jensen, K., Miller, L.A., Byrd, R.J., and Chodorow, M.S., 1982, The EPISTLE Text Critiquing System, *IBM Systems Journal*, 21:3,, 305-326.

Jensen, K., Heidorn, G.E., Miller, L.A., and Ravin, Y., 1983, Parse Fitting and Prose Fixing: Getting Hold on Ill-Formedness, *American Journal of Computational*

Linguistics, 9:3-4, 147-160.

Kucera, H., 1985, Uses of On-Line Lexicons, *Proceedings First Conference of the U.W. Centre for the New Oxford English Dictionary: Information in Data*, University of Waterloo, 7-10.

Salton, G., 1975a, A Theory of Indexing, *Regional Conference Series in Applied Mathematics*, No. 18, Society of Industrial and Applied Mathematics, Philadelphia, PA.

Salton, G., Yang, C.S., and Yu, C., 1975b, A Theory of Term Importance in Automatic Text Analysis, *Journal of the ASIS*, 26:1, 33-44.

Walker, D.E., 1987, Knowledge Resource Tools for Analyzing Large Text Files, in *Machine Translation: Theoretical and Methodological Issues*, Sergei Nirenburg, editor, Cambridge University Press, Cambridge, England, 247-261.

Figure 1. Typical Book Index Entries

Document 659
.T
A Highly Associative Document Retrieval System
.W

This paper describes a document retrieval system implemented with a subset of the medical literature. With the exception of the development of a negative dictionary, all system operations are completely automatic. Introduced are methods for computation of term-term association factors, indexing, assignment of term-document relevance values, and computations for recall and relevance. High weights are provided for low-frequency terms, and retrieval is performed directly from highly connected term-document files without elaboration. Recall and relevance are based on quantitative internal system computations, and results are compared with user evaluations.

Figure 2. Typical Document Abstract

```
DECL   PP      PREP     "with"
              DET      ADJ*       "the"
              NOUN*    "exception"
              PP       PREP       "of"
                       DET        ADJ*          "the"
                       NOUN*      "development"
                       PP         PREP          "of"
                                  DET           ADJ*          "a"
                                  AJP           ADJ*          "negative"
                                  NOUN*         "dictionary"
                                  PUNC          ","
       NP      QUANT    ADJ*       "all"
              NP       NOUN*      "system"
              NOUN*    "operations"
       VERB*  "are"
       AJP     AVP      ADV*       "completely"
              ADJ*     "automatic"
       PUNC   "."
```

Figure 3. Typical Output of Syntactic Analysis Program for One Sentence

assignment computation	negative dictionary
association assignment	quantitative computations
association computations	recall computations*
association factors	relevance values*
association indexing	retrieval system (T)
associative retrieval (T)*	subset implemented
associative system (T)	system computations
computations computation	system implemented
computation methods	system operations
connected file	term-document files
development exception	term-document relevance
dictionary development	term-document relevance values
document retrieval (T,2)*	term-document values *
document retrieval system (2)	term-term-assingment
document system (T,2)	term-term association *
elaboration files	term-term association factors
factors computation	term-term computation
indexing computation	term-term factors
internal computation	term-term indexing
literature subset	user evaluation *
low-frequency terms	values assignment
medical literature	

Figure 4. Phrases generated for Document 659
(T = title; 2 = occurrence frequency of 2; * = manually selected)

1. **Three-Term Phrases**

document retrieval system	
term-term assocaition factor	
term-term relevance values	

2. **Two-Term Phrases** (with Term Frequency greater than 1)

Phrase	Frequency in Document (tf)	Number of Documents for Phrase (out of 1460) (df)
retrieval system	2	125
*document system	2	25
term-term computation	2	1
term-document	2	1
term-term factors	2	1
*term-term indexing	2	5
*document retrieval	2	28
*term-term association	2	2
*term-term assignment	2	2

3. **Two-Term Phrases in Normalized (tf x idf) Weight Order** (df > 1)

Phrase	Weight	Phrase	Weight
term-term assignment	.2128	association factors	.1064
term-term association	.2128	associative system	.1064
term-term indexing	.1832	low frequency terms	.1064
document system	.1313	associative retrieval	.1064
document retrieval	.1276	literature subset	.1064
indexing computation	.1064	term-document files	.1064

Figure 5. Automatic Phrase Indexing for Document 659