Pergamon

Information Processing & Management, Vol. 30, No. 5, pp. 631-645, 1994

0306-4573/94 $6.00 + .00

0306-4573(93)E0011-C

P-15

MACHINE-AIDED INDEXING AT NASA

JUNE P. SILVESTER, MICHAEL T. GENUARDI, and PAUL H. KLINGBIEL
NASA CASI, 800 Elkridge Landing Road, Linthicum Heights, MD 21090-2934, U.S.A.

(*Received* 8 *February* 1993; *accepted in final form* 16 *November* 1993)

Abstract—This report describes the NASA Lexical Dictionary (NLD), a machine-aided indexing system used online at the National Aeronautics and Space Administration's Center for AeroSpace Information (CASI). This system automatically suggests a set of candidate terms from NASA's controlled vocabulary for any designated natural language text input. The system is comprised of a text processor that is based on the computational, nonsyntactic analysis of input text and an extensive knowledge base that serves to recognize and translate text-extracted concepts. The functions of the various NLD system components are described in detail, and production and quality benefits resulting from the implementation of machine-aided indexing at CASI are discussed.

INTRODUCTION

The National Aeronautic and Space Administration's (NASA's) machine-aided indexing (MAI) system is fully operational and cost effective. It is a third generation of a system designed by Paul H. Klingbiel for the Defense Technical Information Center (DTIC). This article describes the NASA system, which was developed as part of a concentrated effort to speed up the indexing of scientific and technical reports and cut costs. The system functions within normal NASA time constraints and workloads and is used in conjunction with electronic input processing. Although NASA has conducted a number of tests to evaluate its MAI system, measures were restricted to those that would not slow production. The best proof of the success of MAI is that indexers handle more work than ever before, they like MAI, and there has been no adverse effect on retrieval, as evidenced by user or retrieval analysts' complaints.

NASA's system can be used for batch processing. More often it is used interactively during online document processing. NASA's MAI is designed as a tool for indexers, and all output is expected to be reviewed. To make processing fast enough for an online system, NASA replaced DTIC's method of machine selection of phrases and expanded the knowledge base (KB), which is used as a kind of conceptual network. This is described at greater length in this article.

DTIC's phrase delineation method, which is the method that NASA originally employed, is based on discovering noun phrases. This system uses a recognition dictionary to assign syntax to each word encountered in text and a Machine Phrase Selection (MAPS) program to string words together according to specified grammar rules (Silvester *et al.*, 1993a). The object of MAPS is to identify the noun phrases that exist in natural language text.

The first DTIC system required that the entire phrase identified by MAPS be located as a key to an entry in their Use Reference file, which they called the Natural Language Data Base (NLDB). This file became very large and cumbersome, but Klingbiel discovered that its contents could be reduced greatly. DTIC's current system has replaced the NLDB with a more condensed file called the Lexical Dictionary (Klingbiel, 1985), which was the pattern for NASA's original knowledge base.

In the syntactic system described earlier, any word that is not likely to be part of a noun phrase becomes a stopword and is considered to be without indexable content. This includes verbs, since verbs are not part of a noun phrase. For DTIC's system, about

Correspondence should be addressed to NASA CASI, 800 Elkridge Landing Road, Linthicum Heights, MD 21090-2934.

50% of the words occurring in text are classified as stopwords, and NASA's original MAI system had more than 76,000 such words.

NASA's present system is not based on grammatical parsing. It is semantically based. Bonnie Jean Dorr describes several arguments for choosing a semantic-based design over a syntactic one (Dorr, 1988), such as (1) the large number of rules required for a syntactic-based system to handle different meanings of context sensitive words, (2) the enormous amount of information needed to disambiguate words, and (3) the attention of syntactic systems to form rather than content. NASA's primary reason for using a semantic-based system was processing speed for online MAI. NASA's present system is based on the co-occurrences in parts of a sentence of what Alan Melby calls domain-specific terminology (Melby, 1990). This refers to words and phrases that are not broad in their meanings but that have (or suggest) domain-specific, semantically unambiguous, indexable concepts. These text words and phrases are matched against a list of text words and phrases that are generally synonymous to NASA's thesaurus terms. They are linked to NASA's thesaurus terms in the knowledge base. Susan Artandi describes this as an "inclusion list" approach (Artandi, 1976). Such a list can accomplish synonym control by switching from text words to specified index terms, and the input text is not limited to noun phrases. It may contain verbs, adverbs, and even nongrammatical word combinations that contribute to the recognition of indexable concepts. When grammatical parsing was abandoned as the way in which to identify indexable concepts, it was possible to capture indexable concepts previously discarded because they were not expressed in noun phrases. This improved the quality of the final output. The semantic-based system significantly reduced the text processing time. It was also possible to reduce the stopword list to about 250 statistically selected words. These 250 stopwords and punctuation are used to break the text into word strings.

The current online NASA system is constructed as a three-component system (see Fig. 1):

- The knowledge base (KB), which is the dataset used for MAI;
- Application programs; and
- Access-2, a modular program that
  - Constructs search keys by concatenating words within established boundaries;
  - Looks up the search keys in the KB; and
  - Returns the candidate NASA Thesaurus posting terms and any other reports to the application program for output to the user.

These components are described in greater detail next.

THE KNOWLEDGE BASE

The lists that indicate which thesaurus terms to use for any given input have been referred to by several names during the past 10 years. In 1982 at NASA, the dataset was called the NASA Lexical Dictionary (NLD), after the corresponding DTIC dataset. Soon the NASA Lexical Dictionary (NLD) became a system with three Use Reference files:

- One for mapping DTIC Thesaurus descriptors to NASA's—a process that we call Subject Switching (Silvester *et al.*, 1984; Silvester & Klingbiel, 1993);
- A second dataset for DOE to NASA Subject Switching; and
- A third dataset for mapping natural language to NASA thesaurus terms, referred to (after the separation of the original dataset into three) as the Phrase Matching file.

The Phrase Matching file contained about 66,000 entries when NASA began to use it operationally. It has since developed into a knowledge base that, as of July 1993, contained more than 115,000 entries. The ultimate goal is to continue to expand the KB and to unify these three datasets insofar as it is feasible and compatible with natural language MAI. The KB could reach 250,000 entries before significant stabilization occurs. The growth of the KB is shown in Fig. 2. This, as we have stated, is the dataset that provides

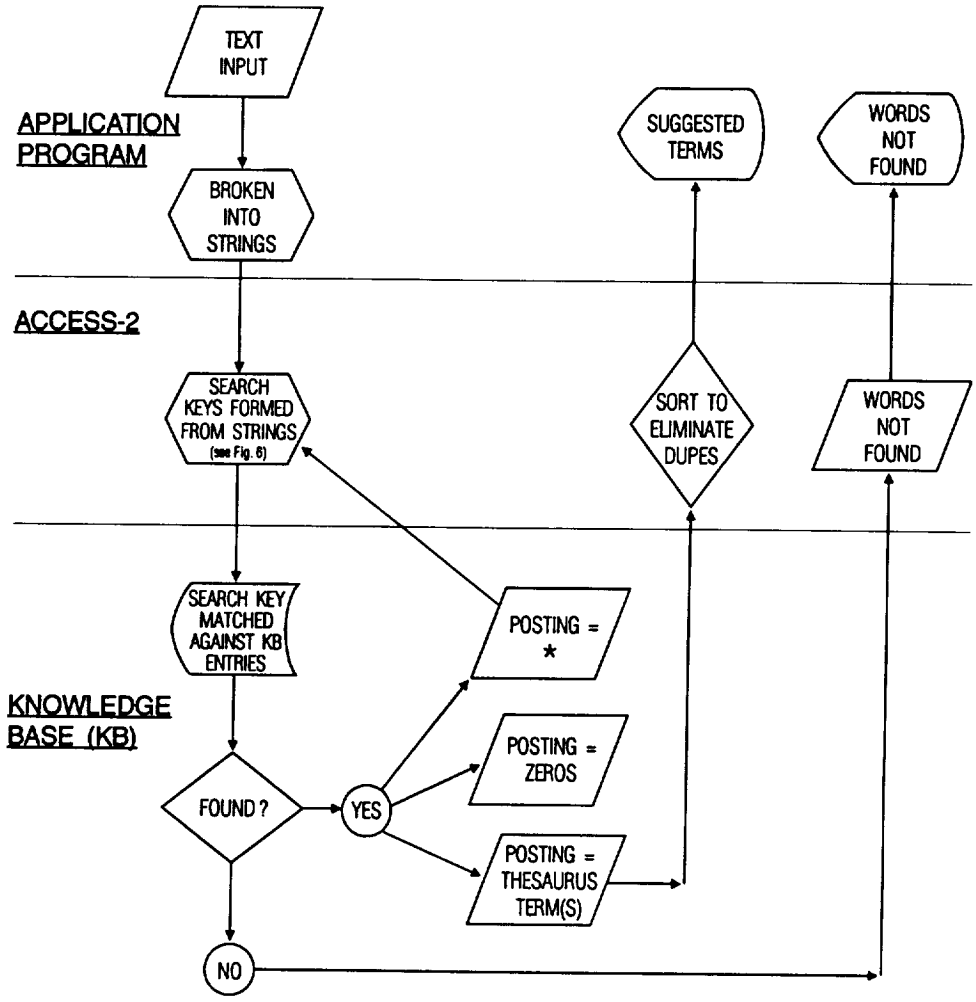


Fig. 1. Overview of NASA's online machine aided indexing system.

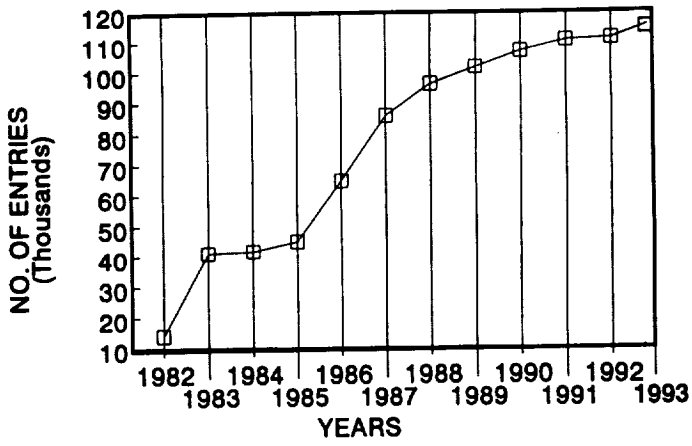
the translations from natural language to NASA's controlled vocabulary (thesaurus terms). Two fields are essential:

- The Key field of each record, which is unique and serves as the computer address to the entry in the KB, and
- The Posting Terms field.

The unique Key field consists of one of the following:

- Any word followed by a semicolon and three nines. Nines are used because they sort last in NASA's IBM-4381 mainframes, on which MAI is processed.* (A single word followed by 999 must sort last because that entry is the default lookup, which is of interest only when other combinations beginning with the initial word are not found. The word combinations beginning with the same initial word are searched sequentially in the computer's sort order. Sort order on the IBM mainframe begins with spaces and symbols, followed by alphas, and ending with numbers 0 through 9.)
- A combination of two or more words separated by semicolons.

*At present, indexers have access to the mainframe from a 3270-type terminal.



YEAR	ENTRIES	YEAR	ENTRIES
1982	14,000	1988	96,300
1983	41,000	1989	101,800
1984	41,700	1990	107,000
1985	45,000	1991	110,700
1986	64,800	1992	111,500
1987	86,000	1993	115,400

1983-1984 SHOWS EFFECT OF REDUCTION IN RESOURCES
1991-1992 ALSO SHOWS EFFECT OF REDUCTION IN RESOURCES FOR THIS PARTICULAR TASK

Fig. 2. Knowledge base growth.

- A combination of two or more words separated by semicolons that becomes a unique combination of characters by the addition of ";999".

See the examples in Fig. 3.

Posting terms are known by a variety of names. In this document we have tried to restrict our terminology for that concept to either posting terms or NASA thesaurus terms.

The Posting Terms field contains

- One or more thesaurus terms that are equivalent in meaning to the key; or
- Two zeros (00); or
- An asterisk (*).

See the examples in Fig. 3.

The Key field and Posting Term field are a more robust rewrite system than that of Use References in a standard thesaurus. The usual Use reference types (with the addition of semicolons and the 999 flag) occur as keys, but a new and powerful concept is added—a rewrite to 00 in the Posting Term field. Linguistically, this deletion is a zeroing rule.

As a conceptual network, the knowledge base contains not only entries that map natural language words and phrases to controlled vocabulary terms, but also entries that represent decisions regarding the relevancy of particular concepts (Genuardi, 1990). For example, within the aeronautics domain, the concept AIRCRAFT is much too broad in meaning to be a useful indexing term for most instances in which the word *aircraft* appears in text. In this case, specific entries in the knowledge base would initiate a search for a multiword semantic unit such as A-320 AIRCRAFT, which describes the specific vehicle in question; or AIRCRAFT STABILITY, AIRCRAFT CONSTRUCTION MATERIALS, or AIRCRAFT CONFIGURATIONS, which indicate the particular aeronautical aspect of interest. Other entries in the knowledge base serve to disambiguate certain words such as *matrices*, which might refer to either mathematical matrices or material matrices.

Keys	Posting Terms
INDIRECTLY;999	00
INDIUM;OXIDE	*
INDIUM;OXIDE;COATING	COATINGS, INDIUM COMPOUNDS
INDIUM;OXIDE;COATINGS	COATINGS, INDIUM COMPOUNDS
INDIUM;OXIDE;999	INDIUM COMPOUNDS, METAL OXIDES
INDIUM;999	INDIUM
LIQUID;PHASE	*
LIQUID;PHASE;EPITAXY	LIQUID PHASE EPITAXY
LIQUID;PHASE;999	LIQUID PHASES
LIQUID;ROCKET	*
LIQUID;ROCKET;PROPELLANT	LIQUID ROCKET PROPELLANTS
LIQUID;ROCKET;PROPELLANTS	LIQUID ROCKET PROPELLANTS
LIQUID;ROCKET;999	00
LIQUID;999	00

Note that in the Key field there are no spaces, and in the Posting Term field multiple thesaurus terms are separated by comas with spaces occurring only between words in a multiword term. In a file as large as the NASA knowledge base, this practice not only saves storage space, but more importantly, it saves computer reading time. Any procedure that saves computer processing time is vital to an online MAI system.

Fig. 3. Examples of keys and posting terms.

The process of identifying KB entries is similar to the one described by N. Vleduts-Stokolov for specifying "concept codes" from word co-occurrences in the BIOSIS database (Vleduts-Stokolov, 1982). The current method of selecting KB entries is based on a statistical analysis of the single- and multiword phrases that occur in large volumes of text (Genuardi, 1990). These phrases occur in text that (1) resides in the NASA database, (2) is indexed to a targeted thesaurus term, and (3) contains the candidate words or phrases with relative frequency. The phrases selected are phrases that would be used by the NASA MAI process as search keys.

In general, the procedures selected for an MAI system's initial phrase delineation and analysis define what kinds of information need to be represented in knowledge base entries and how large an operational file will need to be. For example, the use of word stemming or phrase normalization could reduce the number of required entries. Likewise, the strategies used for disambiguating words and for analyzing relevancy can define the level of complexity required for knowledge representation and ultimately may dictate the kind of data structure that is used. In the particular case of the NASA knowledge base, when the trade-offs were considered, it was decided to keep all rules as simple as possible to keep

the system's online response time as short as possible. By rules, we mean "if... then" statements. For example, if "In-102" is encountered in the title or abstract, then provide the Thesaurus term "INDIUM ISOTOPES" as a suggested term for indexer review. Or, if a word is hyphenated, then look in the knowledge base for the hyphenated form; if it is found, then read the Posting Term field; else (if it is not found) drop the hyphen and treat the hyphenated word as two separate words. Most rules in the NASA MAI system consist of rules that specify (1) if the search key is found and the Posting Term field contains NASA Thesaurus terms, then suggest the NASA Thesaurus term(s) for review by the indexer; or (2) if the search key is found and the Posting Term field contains an asterisk, then add the next word in the five-word array to the search key and look up the new search key; or (3) if the search key is found and the Posting Term field contains two zeros, then no translation to NASA thesaurus terms is wanted for that word or word combination. Some MAI systems have more numerous rules that will examine instances of capitalization of words in the key or look for specific words in close proximity to a word in the key as part of the "if" statement (M. M. J. Hlava, personal communication, October 13, 1992). For example, if the word *titanic* occurs, and if it begins with an upper-case T, and if the word *ship* occurs within four words of *Titanic*, then return the term *U.S.S. Titanic* for indexer review. The NASA system designers chose to forego such details in the interest of minimizing the reading and writing required of the computer and thereby maximizing the speed of processing.

APPLICATION PROGRAMS

Each new use for MAI requires an application program that

- Identifies the source of the text to be processed;
- Delineates word strings found in natural language text by establishing boundaries or parameters;
- Removes parentheses unless they are embedded (i.e., unless there are characters on both sides of a parenthesis);
- Checks any word with an embedded hyphen (-) or virgule (/) against the first words in the KB keys and, if not found, drops the embedded symbol and treats the word as two words;
- Calls Access-2;
- Receives the MAI output for that particular application; and
- Writes out the suggested terms plus any reports specified.

Input can be any pertinent text. Usually it consists of titles and abstracts; however, it can be the subject terms or descriptors from another organization's controlled vocabulary, material from indexed or unindexed documents, the first and last sentence of designated paragraphs, an executive summary, or any other text specified by the user and identified by the application program. The program uses a table of about 250 statistically selected stopwords (see Fig. 4) and thought-ending punctuation such as colons, semicolons, and periods. As punctuation or stopwords are encountered, the string ends and it is ready for processing by Access-2. Note that the stopword list does not contain either *a* or *the*. To include these as stopwords would preclude the ability to recognize valid thesaurus terms and Use references that contain these words, such as BOMARC A MISSILE, A STARS, VITAMIN A (Use RETINENE), OVER-THE-HORIZON RADAR, and LOGISTICS OVER THE SHORE (LOTS) CARRIER.

ACCESS-2

Access-2, which is a modular program, never acts by itself. It is always called by an application program. Access-2 was designed to replace the syntactic analysis which characterized the original Machine Phrase Selection program. In addition, it was designed to shorten the overall machine processing time by minimizing I/O (the transfer of data

ABOUT	DEMONSTRATED	I.E	PARTICULAR	SUGGESTED
ABOVE	DESCRIBE	IF	PAST	SUITABLE
ACCOUNT	DESCRIBED	IMPLEMENTATION	PERFORMED	SUMMARY
ACHIEVED	DESCRIBES	IMPORTANCE	POSSIBLE	TAKEN
ACROSS	DESIGNED	IMPORTANT	PREDICT	TESTED
ADDITIONAL	DETAILED	IMPROVE	PREDICTED	THAN
AFTER	DETERMINE	INCLUDE	PRELIMINARY	THAT
ALLOW	DETERMINED	INCLUDED	PRESENCE	THEIR
ALLOWS	DETERMINING	INCLUDES	PRESENT	THEM
ALONG	DEVELOP	INCLUDING	PRESENTED	THEN
ALSO	DEVELOPED	INCREASE	PRESENTS	THERE
ALTHOUGH	DIFFERENT	INCREASED	PREVIOUS	THESE
AMONG	DIRECTLY	INCREASES	PREVIOUSLY	THEY
AN	DISCUSSED	INDICATE	PRODUCE	THIS
ANY	DOES	INDIVIDUAL	PRODUCED	THOSE
APPROPRIATE	DUE	INTEREST	PROPOSED	THROUGH
APPROXIMATELY	DURING	INTO	PROVIDE	THUS
ARBITRARY	E.G	INTRODUCED	PROVIDED	TOGETHER
ARE	EACH	INVESTIGATE	PROVIDES	TOWARD
AROUND	EFFICIENT	INVESTIGATED	PROVIDING	TYPES
AS	EFFORTS	INVOLVED	RECENT	TYPICAL
ASPECTS	EITHER	INVOLVING	RELATED	UNDERSTANDING
ASSOCIATED	EMPHASIS	IS	RELATIVELY	UNIQUE
ASSUMED	EMPLOYED	ISSUES	REPORTED	UP
AVAILABLE	ESPECIALLY	IT	REQUIRED	UPON
BASIS	ESTABLISHED	ITS	REQUIRES	USED
BECAUSE	EVALUATE	KNOWN	RESPECT	USEFUL
BEEN	EVALUATED	LESS	RESULT	USES
BEING	EXAMINED	MADE	RESULTING	USING
BEST	EXAMPLE	MAJOR	RESULTS	VARIETY
BETTER	EXAMPLES	MAKE	REVIEWED	VARIOUS
BOTH	EXISTING	MAY	RTOP	VERSION
BUT	EXPECTED	MEANS	SAME	VIA
CAN	EXPERIMENTALLY	MORE	SELECTED	WAS
CARRIED	FEW	MOST	SEVERAL	WE
CAUSED	FOUND	MUCH	SHOULD	WERE
CERTAIN	FULLY	MUST	SHOW	WHEN
CHARACTERIZED	FUNDAMENTAL	NECESSARY	SHOWED	WHERE
COMPARED	FURTHER	NEED	SHOWN	WHICH
COMPLETE	GIVEN	NEEDED	SHOWS	WHILE
CONSIDERATION	GOOD	NOT	SIGNIFICANT	WHOSE
CONSIDERED	GREATER	OBJECTIVE	SIGNIFICANTLY	WILL
CONSISTS	HAD	OBSERVED	SINCE	WITH
CONTAINING	HAS	OBTAIN	SOME	WITHIN
CONTAINS	HAVE	OBTAINED	STATUS	WITHOUT
CONVENTIONAL	HAVING	OCCUR	STUDIED	WOULD
CORRESPONDING	HERE	OTHER	STUDIES	YEARS
COULD	HOW	OUR	STUDY	
DEFINED	HOWEVER	OVERALL	SUB	
DEMONSTRATE	IDENTIFIED	PART	SUCH	

Fig. 4. List of stopwords used with ACCESS-2.

between the central processing unit and the KB or any other file). This is achieved primarily by eliminating lookups to determine parts of speech for each word encountered and lookups to find the appropriate grammar rules to follow—that is, by eliminating parsing. Speed is also achieved by keeping rules as simple as possible and reducing the number of unneeded empty spaces in a record that must be read by the computer. The fewer times that files must be accessed, and the fewer times that information must be written to a file or to a monitor, the shorter the response time. With the original MAI system, including the Recognition Dictionary, the Machine Phrase Selection program, and Access-1, a title and a 250- to 300-word abstract could be processed in approximately 1.5 minutes. However, syntactic analysis was found to be unnecessary for quality performance of the system. With Access-2 the response time averages 6 to 7 seconds for the same amount of text.

When strings have been delineated, Access-2 identifies semantic units contained within the string. The semantic unit in the NASA system is normally limited to a maximum of five words to ensure grammatically correct word associations without parsing; however, the system can handle longer units if the words are consecutive. Search keys of fewer than five words must be created from within a five-word segment of the machine-selected string.

This five-word proximity limit was established empirically and represents the best trade-off between identifying the most semantic units while limiting the risk of inappropriate word concatenations.

When Access-2 accepts each word in an input string from the application program, and before it begins to identify semantic units, it places each word into its own array cell.

Identifying semantic units is done as follows:

- The computer-selected strings are examined, from left to right, in five word segments, beginning with word one and word two. The first word of every word combination is checked against the KB to see if it exists. If it does not, the word is stored in a list of "Words Not Found As First Word in a Key" and printed out for indexer review.
- If word one followed by word two is found in the KB as a key to an entry, the posting term field of that entry, which contains the equivalent NASA thesaurus term(s), is read. There are three possibilities (see Figs. 1 and 3):
  - The posting term field may contain two zeros (00), which will generate no NASA thesaurus term or terms; or
  - The posting term field may contain one or more thesaurus terms that are equivalent or slightly broader in meaning to the key and that will be provided to the indexer as suggested indexing terms; or
  - The posting term field will contain an asterisk (*), which causes the program to look for an additional word (within the five-word segment) that, when added to the two previous words, will match the key of another record.
- If word one followed by word two has an asterisk in the posting term field, and this combination followed by word three, or four, or five does not find a matching key in the knowledge base, then the program adds 999 (which sorts last) in place of the final word and tries that combination as a key. If that is not found, the final word in the candidate key is dropped and replaced with 999. This procedure is repeated, if necessary, until the key is reduced to the first word and 999.
- If word one followed by word two is not found in the knowledge base, then word one is looked up with word three.
- If word one has been tried with each other word in the five-word segment and no key leading to a thesaurus term is found, the computer looks up word one followed by 999 to see if a thesaurus term is provided for the single word. This is possible for single words that represent specific indexable concepts.
- When the process has used or rejected word one, the five-word segment is again measured off, beginning with word two.
- Once a word is found as part of a KB entry, it is "poisoned"—that is, it is stored with a flag appended to it until the processing has passed that word. A poisoned or flagged word may not be used again unless an unpoisoned word is added to it. (See the following example of AERODYNAMIC CONFIGURATIONS.)
- If word one and word two are found in the KB and word three is *and* or *or*, the last word in the key is dropped and the first word is combined with the word that follows *and* or *or* to form a new search key.

For example, consider the following five-word segment:

"aerodynamic configurations and properties that..."

Look up search key "AERODYNAMIC;CONFIGURATIONS". Find the thesaurus term "AERODYNAMIC CONFIGURATIONS".

The program "poisons" (or flags) "AERODYNAMIC" and "CONFIGURATIONS". These words may (now) be used again only if combined with a new word.

The next word is "AND". The program drops "AND" and concatenates the word "AERODYNAMIC" with the next word, which is "PROPERTIES" and which has not yet been poisoned or combined with "AERODYNAMIC."

Look up the search key "AERODYNAMIC PROPERTIES". Find the thesaurus term "AERODYNAMIC CHARACTERISTICS", poison the word "PROPERTIES", and conclude the processing for word 1, "AERODYNAMIC".

The next five-word segment is counted off, beginning with the word after the conjunction, and the process begins again with "PROPERTIES" (now poisoned) as word 1.

A more complete example of processing text with Access-2 is illustrated in the following example. Given the following title and sentences from an abstract of a document:

> Helicopter Noise
>
> Acoustic data for a 40 percent model MBB BO-105 helicopter main rotor were obtained from wind tunnel testing and scaled to equivalent actual flyover cases. It is shown that during descent the dominant noise is caused by impulsive blade-vortex interaction (BVI) noise. In level flight and mild climb BVI activity is absent; the dominant noise is caused by blade-turbulent wake interaction.

KB file entries needed to process the sample input, and some related KB entries have been extracted and are listed in Fig. 5.

Word strings are delineated by means of stopwords or any thought-ending punctuation such as a period, colon, or semicolon. This delineation process produces the following word strings from the foregoing title and abstract:

1. helicopter noise
2. acoustic data for a 40 percent model MBB BO-105 helicopter main rotor
3. from wind tunnel testing and scaled to equivalent actual flyover cases
4. descent the dominant noise
5. by impulsive blade-vortex interaction (BVI) noise
6. in level flight and mild climb BVI activity
7. absent
8. the dominant noise
9. by blade-turbulent wake interaction

Key	Posting Term
ACOUSTIC;DATA	*
ACOUSTIC;DATA;CAPSULE	ACOUSTIC PROPERTIES
ACOUSTIC;DATA;999	ACOUSTIC PROPERTIES
BLADE-VORTEX;INTERACTION	BLADE-VORTEX INTERACTION
BLADE-VORTEX;TURBINE	TURBINE BLADES
BLADE;VORTEX	*
BLADE;VORTEX;INTERACTION	BLADE-VORTEX INTERACTION
BLADE;999	00
BO-105;HELICOPTER	BO-105 HELICOPTER
BO-105;HELICOPTERS	BO-105 HELICOPTER
CLIMB;999	CLIMBING FLIGHT
DATA;999	00
DESCENT;999	DESCENT
HELICOPTER;NOISE	AEROACOUSTICS,AERODYNAMIC NOISE,AIRCRAFT NOISE
HELICOPTER;ROTOR	*
HELICOPTER;ROTOR;NOISE	AEROACOUSTICS,AERODYNAMIC NOISE,AIRCRAFT NOISE
HELICOPTER;ROTOR;999	ROTARY WINGS
HELICOPTER;ROTORS	ROTARY WINGS
TURBULENT;WAKE	TURBULENT WAKES
TURBULENT;WAKES	TURBULENT WAKES
TURBULENT;999	TURBULENCE
WIND;TUNNEL	*
WIND;TUNNEL;TEST	WIND TUNNEL TESTS
WIND;TUNNEL;TESTING	WIND TUNNEL TESTS
WIND;TUNNEL;TESTS	WIND TUNNEL TESTS
WIND;TUNNEL;999	WIND TUNNELS
WIND;TUNNELS;999	WIND TUNNELS

Fig. 5. KB file entries needed to process the sample input.

In the following example of Access-2 processing, references are made to the input array and the KB entries just shown. For a flow chart showing the processing logic, see Fig. 6, which explains how search keys are formed from word strings. It may also be helpful to refer again to Fig. 1, which provides an overview of the system.

Processing Descriptions and Outcomes

Mark off five-word array in title. Outcome: Only two words exist; therefore, the array is "Helicopter Noise."

Look up search key "HELICOPTER;NOISE" in KB. Outcome: Key found. Posting term(s) "AEROACOUSTICS,AERODYNAMIC NOISE,AIRCRAFT NOISE" returned.

No more words exist in the title. Move to the first string in the abstract. Outcome: The first MAI-selected string in the abstract is "Acoustic data for a 40 percent model MBB BO-105 helicopter main rotor."

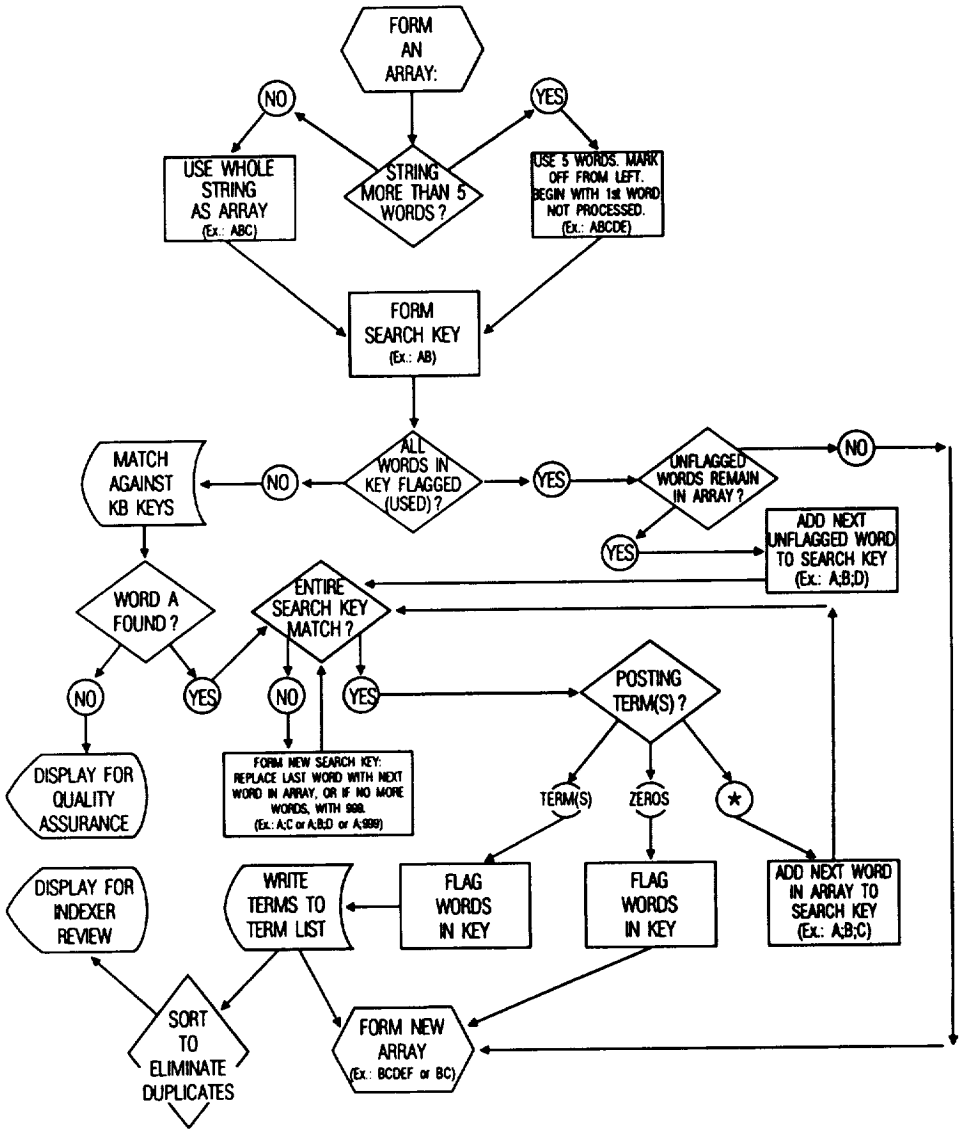


Fig. 6. Process for forming search keys from word strings.

Procedure

Mark off the first five-word array in the string; concatenate word 1 and word 2 to form a search key (in the sample string this would be ACOUSTIC;DATA), and look up search key in the KB. Outcome: Key found. Posting term "ACOUSTIC PROPERTIES" returned.

If the key leads to a posting term(s) or 00, poison (flag) the words in the key (e.g., the words "ACOUSTIC" and "DATA") and end processing for word 1 (e.g., ACOUSTIC). Outcome: A poisoned word may not be used again unless it is combined in a search key with an unpoisoned word.

Move one word to the right in the string and mark off a new five-word array (e.g., DATA FOR A 40 PERCENT). Concatenate the new array's word 1 with the new array's word 2, and look up the search key (e.g., DATA;FOR). Outcome: Key not found.

If key is not found, look up the next search search key(s) in this array; that is, concatenate words 1 and 3, 1 and 4, 1 and 5 (e.g., "DATA;A," "DATA;40," "DATA; PERCENT"). Outcome: Keys not found.

If an asterisk is found in the posting term field, then the key must have an additional word or 999 in order to translate. For example, if words 1 and 3 lead to an asterisk, look next for 1, 3, and 4; 1, 3, and 5; and finally, for 1, 3, and 999. Whenever a key is not found and untried words remain in the five-word array, continue processing combinations that begin with word 1.

End the processing for word 1 whenever the KB provides output for a key, or word 1 has been tried unsuccessfully with words 2, 3, 4, and 5.

End Procedure

Mark off the next five-word array in the string and repeat the *Procedure* described. The remaining arrays for the first string are as follows:

"FOR A 40 PERCENT MODEL"

"A 40 PERCENT MODEL MBB"

"4O PERCENT MODEL MBB BO-105"

"PERCENT MODEL MBB BO-105 HELICOPTER"

"MODEL MBB BO-105 HELICOPTER MAIN"

"MBB BO-105 HELICOPTER MAIN ROTOR"

Outcome: Keys not found.

If fewer than five words remain in the string, accept a smaller segment and follow the same procedures. Only four words remain in the sample string—"BO-105 HELICOPTER MAIN ROTOR". Mark them off and look up the search key of word 1 and 2 (i.e., "BO-105;HELICOPTER"). Outcome: Key found. Posting term "BO-105 HELICOPTERS" returned.

Poison (flag) "BO-105" and "HELICOPTER." Outcome: These words may not be used again without an unpoisoned word.

End processing for "BO-105." This was the initial word of a key that successfully matched a key in the KB and provided a NASA thesaurus term. Outcome: Three words now remain in the string—"HELICOPTER MAIN ROTOR"—and this becomes the new array.

Look up search key "HELICOPTER;MAIN". Outcome: Key not found. (Note: "HELICOPTER" has been poisoned, but it is coupled with "MAIN", which has not been poisoned.)

Look up the next search key: "HELICOPTER;ROTOR" Outcome: Key found. "ROTOR" has not been poisoned. The Posting Term field holds an asterisk (*), which is returned.

An asterisk indicates that another word is needed. There are no more words in the array; therefore add ";999" to the search key and look up "HELICOPTER;ROTOR; 999". Outcome: Key found. Posting term "ROTARY WINGS" returned.

Poison (flag) "ROTOR". Outcome: The word "ROTOR" may not be used again without an added unpoisoned word or words.

Two words now remain in the string—"MAIN ROTOR"—and this becomes the new array. Look up the search key "MAIN ROTOR". "ROTOR" has been poisoned, but "MAIN" has not been part of any key that has been found. Outcome: Key not found.

No more words remain in string. Outcome: End processing for this string.

Repeat process described for the remaining strings. Outcome: No keys that begin with the first word in the first array are found.

The second five-word array in the next string (i.e., "WIND TUNNEL TESTING AND SCALED") illustrates how the KB entries direct the need to concatenate words in an array.

Look up search key "WIND;TUNNEL". Outcome: Key found. Posting term field contains an asterisk (*), which requires the addition of another word from the five-word array.

Add the next word in the array and look up search key "WIND;TUNNEL;TESTING". Outcome: Key found. Posting term "WIND TUNNEL TESTS" returned.

The only other search keys found in the aforementioned strings are DESCENT;999 (the search keys "DESCENT;THE", "DESCENT;DOMINANT", and "DESCENT; NOISE" were not found, and so the final word was replaced with "999"). Outcome: Key found. Posting term "DESCENT" returned.

Look up search key "BLADE-VORTEX;INTERACTION". Outcome: Key found. Posting term "BLADE-VORTEX INTERACTION" returned.

Look up search key "CLIMB;999". Outcome: Key found. Posting term "CLIMBING FLIGHT" returned.

Look up search key "TURBULENT;WAKE". Outcome: Key found. Posting term "TURBULENT WAKES" returned.

In summary, the following terms were suggested:

HELICOPTER NOISE,
AEROACOUSTICS,
AERODYNAMIC NOISE,
AIRCRAFT NOISE,
ACOUSTIC PROPERTIES,
BO-105 HELICOPTERS,
ROTARY WINGS,
WIND TUNNEL TESTS,
DESCENT,
BLADE-VORTEX INTERACTION,
CLIMBING FLIGHT, and
TURBULENT WAKES.

Note that the text that was processed contained several hyphenated words. The application program checks each word with an embedded hyphen or virgule (i.e., a diagonal line, /) against the initial word of the keys in the KB. The compound words BO-105 and BLADE-VORTEX are in the KB, so the hyphens in these words are kept, and the com-

pounds are treated as a single word. However, BLADE-TURBULENT is not found in an initial position in any KB key; therefore, the hyphen between these words is dropped, and the compound is treated as two words.

IMPACT EVALUATION

In NASA's high-pressure production environment, testing the MAI system without slowing the work that must be done has presented a challenge. The number of indexers currently abstracting and indexing is about 40% lower than before the institution of MAI; however, MAI is not the only change responsible for increased productivity. Other indexing aids, such as automatic term validation and online thesaurus access, have been incorporated into the indexing workstation. These features also speed processing. The fact remains that the size of the present indexing staff is about 62% of the pre-MAI staff size and the output of each individual has approximately doubled in the past 10 years.

It was determined in an early test that machine-aided indexing saved an average of 3 minutes per document by reducing the time needed to look up terms in the thesaurus (Silvester *et al.*, 1984). It is reasonable to expect that this time saving is even greater for comparatively new indexers, which now represent about 80% of the staff. It takes time to become familiar with the more than 17,000 terms in the NASA controlled vocabulary. This process is speeded up somewhat by ready access to the thesaurus terms online and by providing good-quality output to indexers through MAI.

Match rate

The first measure of how well the MAI system performed was referred to as the match rate. It describes the percentage of machine-selected phrases that either partially or completely matched keys in the KB. As this percentage rose to nearly 100%, the measure was changed to refer to the percentage of MAI suggested terms that the indexer elected to use. By the second definition, this ratio has grown from 23% to over 50%. In one file up to 79% of the suggested terms were used. Typically, the rate tends to rise gradually as improvements are made to the system.

Capture rate

In November of 1986, NASA instituted another measure referred to as the capture rate. This described the percentage of indexer-assigned terms that were suggested by MAI. The capture rate has been, rather consistently, a few percentage points higher than the match rate.

Consistency factor

A third measure, which we began to use in September 1989, was a consistency (or quality) factor q. This measures the percentage of common terms c found in two lists of terms, one generated automatically and represented by a, and the other terms selected intellectually by the indexer and represented by i. Expressed in another way, q is the ratio of the common terms to the unique terms, where $q = c/(a + i) - c$ (Lustig & Knorz, 1986; Lancaster, 1991).

Table 1 shows the match rate, capture rate, and consistency factors calculated for 1987, 1988, and current estimated performance. The 1987 figures are for a sample of approximately 2500 documents. The 1988 figures were based on a sample of 100 documents, and the 1993 figures are from a survey of the indexers currently using the system. (Available test statistics are limited because of the production environment.)

Other tests done on samples of 100 documents produced the results shown in Table 2. (It was determined early in the project that a sample of 30 documents yielded virtually the same statistics as larger samples for the population of records being studied.) An April 1988 test of records in the Scientific and Technical Aerospace Reports (STAR) series, a September 1989 test of DTIC records, and a June 1992 test of DOE records yielded the statistics shown in Table 2.

Table 1. Measures of match rates, capture rates, and consistency factors

Year	Match rate	Capture rate	Consistency factor
1987	32.4%	36.9%	20.8%
1988	37.0	39.0	23.4
1993	50.0	50.0	33.3

For NASA core literature with good abstracts, approximately 60% of the terms suggested by MAI are acceptable, and they comprise about 50% of the assigned index terms.

BENEFITS

Several benefits result from the use of MAI and the online entry system at NASA:

- MAI-suggested terms are presented online in the correct format and are spelled accurately; therefore, they do not have to be keyed in or manually verified against a thesaurus. An online thesaurus that provides a flexible "display and pick" feature is available.
- Indexer research time is reduced because natural language words, phrases, and acronyms and their technical language/thesaurus equivalents are researched before their addition to the KB, thereby presenting the indexer with expert advice.
- Appropriate, unfamiliar, technical terms may be suggested which the indexer would omit without a prompt from MAI.
- MAI-suggested terms function as a checklist of indexable concepts and increase the consistency of indexing.
- The increased number of indexing terms that often results from using MAI provides additional access points for records.
- Spinoffs from MAI provide aids for the proofreaders, the thesaurus lexicographer, and the retrieval analysts.

The NASA MAI system has also been used as a tool for generating appropriate thesaurus terms for records that were indexed before the thesaurus existed (Silvester *et al.*, 1993b). Terms were provided for more than 400,000 records, and the reports collected by the National Advisory Committee for Aeronautics (NACA) prior to NASA's existence are scheduled for similar treatment. MAI provides a spell-checking feature for proofreaders, identifies new thesaurus terms for lexicographers, and finds cross references for the thesaurus definitions volume.

New uses continue to emerge each year, and the system improves each year. The NASA Center for AeroSpace Information is looking forward to quicker responses and higher rates of matching, capturing, and indexer consistency as well as new, previously untried uses for the NASA MAI system.

Table 2. More measures of match rates, capture rates, and consistency factors

Year	Match rate	Capture rate	Consistency factor
1988	34%	38%	N/A
1989	41	42	26%
1992	40	52	29

REFERENCES

Artandi, S. (1976). Machine indexing: linguistic and semiotic implications. *JASIS*, *27*(4), 235–239.

Dorr, B.J. (1988). *A lexical conceptual approach to generation for machine translation.* Arlington, VA: Office of Naval Research (NTIS No. AD-A197356).

Genuardi, M.T. (1990). Knowledge-based machine indexing from natural language text: knowledge base design, development and maintenance. In H. Czap & W. Nedobity (Eds.), *TKE'90: Terminology and knowledge engineering*. Vol. 1. Proceedings of the Second International Congress on Terminology and Knowledge Engineering (pp. 345–351). Frankfurt, Germany: Indeks Verlag.

Klingbiel, P.H. (1985). Phrase structure rewrite systems in information retrieval. *Information Processing and Management*, *21*(2), 113–126.

Lancaster, F.W. (1991). *Indexing and abstracting in theory and practice* (pp. 60–85). Champaign: University of Illinois.

Lustig, G., & Knorz, G. (1986). *AIR/PHYS pilot application project: pilot application of automatic indexing and improved retrieval methods using the PHYS data base (1-30).* Karlsruhe, Germany: Fachinformationszentrum, Energie Physik Mathematik GmbH.

Melby, Alan K. (1990). Benefits and limitations of formal systems in technical writing. In H. Czap & W. Nedobity (Eds.), *TKE'90: Terminology and knowledge engineering*. Vol. 1. Proceedings of the Second International Congress on Terminology and Knowledge Engineering (pp. 24–25). Frankfurt, Germany: Indeks Verlag.

Silvester, J.P., Newton, R., & Klingbiel, P.H. (1984). *An operational system for subject switching between controlled vocabularies: a computational linguistics approach* (NASA Contractor Report No. 3838). Washington, DC: National Aeronautics and Space Administration (NTIS No. N85-11903).

Silvester, J.P., Genuardi, M.T., & Klingbiel, P.H. (1993a). *Machine aided indexing from natural language text* (NASA Contractor Report No. 4512). Washington, DC: National Aeronautics and Space Administration (NTIS No. 93N-26901).

Silvester, J.P., Genuardi, M.I., & Klingbiel, P.H. (1993b). *NASA's online machine aided indexing system* (NASA Contractor Report No. 4518). Washington, D.C.: National Aeronautics and Space Administration (NTIS No. N94-15009).

Silvester, J.P., & Klingbiel, P.H. (1993). An operational system for subject switching between controlled vocabularies. *Information Processing and Management*, *29*(1), 47–59.

Vleduts-Stokolov, N. (1982). An automatic support to indexing a life sciences data base. *Information Processing & Management*, *18*(6), 313–321.