04

ANNUAL REPORT: AUTOMATIC INFORMATIVE ABSTRACTING AND EXTRACTING

L. L. Earl, et al

Lockheed Missiles and Space Company
Palo Alto, California

February 1972

DK

STATENS
INSTITUT
FÖR
FÖRETAGS-
UTVECKLING

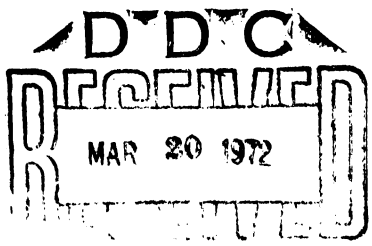
SEE
AD
721066

IVA

ANNUAL REPORT:
AUTOMATIC INFORMATIVE ABSTRACTING AND EXTRACTING

LMSC-D246461 February 1972

Annual Progress Report
Office of Naval Research
Contract N00014-70-C-0239



Information Sciences
Lockheed Palo Alto Research Laboratory
LOCKHEED MISSILES & SPACE COMPANY, INC.
A Subsidiary of Lockheed Aircraft Corporation
Palo Alto, California 94304

PRÉCIS

RESEARCH PROGRESS REPORT

Title: "Annual Report: Automatic Informative Abstracting and Extracting," Annual Progress Report, Office of Naval Research, Contract N00014-70-C-0239.

Authors: L. L. Earl, O. Firschein, and M. A. Fischler

Background: This investigation is concerned with the development of automatic indexing, abstracting, and extracting systems. Basic investigations in English morphology, phonetics, and syntax have been pursued as necessary means to this end. Experimental indexing and extracting systems have been developed. At this time, investigation of the use of syntax in indexing and of descriptive representation of pictorial data is continuing.

Condensed Report Contents: Part I of this report describes a continuing effort in the development of tools for making syntactic and semantic distinctions of potential use in automatic indexing and extracting. One of the tools is a program for syntactic analysis of English; the other is a dictionary of English word government patterns. A multi-level parser PHRASE is described, in which the syntactic structure is built up in four stages, with ambiguities at each stage resolved to yield but one structure for a given sentence. The resultant structure at each level is designed to be useful in its own right, and also to form the basis for the analyses at the next higher level. The nature of the rules for identifying structures and resolving ambiguities is discussed for all four levels of analysis, and examples of the level 1 and 2 analyses, which have been implemented, are given. The nature of word government is discussed and also its usefulness in making semantic and syntactic distinctions. Appendixes give government tables compiled in the last few years.

Part II of the report deals with the three main problems that arise in the storage and retrieval of picture descriptions: (1) the acquisition of meaning from natural language descriptions, (2) the symbolic representation of the meaning, and (3) the organization of the data base of descriptions to allow efficient retrieval of descriptions in response to queries. In the natural language area, the problem of ambiguity is discussed, and a sample parsing of descriptions using the PHRASE parser is presented. The conceptual classes and picture primitives to be used in representing the meaning of descriptions are treated in some detail, concentrating on natural language expressions for "location." In the data base organization area, it is noted that the complexity of description can be reduced by providing the system with "world knowledge" concerning relationships, and a two-dimensional map is suggested for this purpose.

For Further Information: The complete report is available in the major Navy technical libraries and can be obtained from the Defense Documentation Center. A few copies are available for distribution by the authors.

FOREWORD

This report marks the completion of the eighth year in which the Office of Naval Research has contributed support to research in the Information Sciences at the Lockheed Palo Alto Research Laboratory of the Lockheed Missiles & Space Company, Inc. During the first year of the program, a major part of the effort went into establishment of a word-data base. The English Work Speculum, which has been distributed to ONR program participants, illustrates the nature of this data base. In the second and third years, this data base was exploited in the development of a computer program for the automatic assignment of parts of speech to English words. Also during these years, it was demonstrated how an English/Russian phrase data base can be used to develop a technique for obtaining English indexes from untranslated Russian text.

In the third and fourth years, a new data base of sentences with assigned parts of speech was created for investigation of the abstracting and extracting process. Also begun during the third and fourth years were experiments in the compilation of a "sentence dictionary" of syntactic types and compilation of English syntactic word government tables. These activities were continued in the fifth year, along with development of a parsing program, the initiation of some extracting experiments on some technical text, and an experiment in automatic indexing of a medical book. In the sixth year, the "sentence dictionary" experiment was concluded, the extracting experiment was completed, a frequency-syntax method of indexing was conceived and tested, and the concept of English syntactic word government was expanded while compilation of the tables continued.

In the seventh year, compilation of the word government tables was temporarily halted while effort was concentrated in two main areas. First, the scope of the parsing program was extended, preparatory to eventual additional indexing experiments using

syntax in conjunction with frequency and word government criteria. Second, a study in describing and abstracting pictorial structures was undertaken. This year, the extensions to the parsing program were completed and tested, and a plan for a complete four-level parsing system was conceived and described, with the level of descriptive detail differing, of course, according to the extent of current implementation. Also during the year, compilation of English syntactic word government was resumed, in a somewhat augmented form. Finally, a series of experiments involving human subjects describing aerial photographs was completed and the results analyzed, particularly for the "metadescriptive" information in the descriptions and for derivation of canonical forms that can be used to represent the content of the descriptions.

Part I of this report is concerned with the development and uses of the syntactic analyzer and with the concept of English word government. Part II describes the investigations in describing and abstracting pictorial structures.

The group at Lockheed takes this opportunity to express their thanks for the continuing support and encouragement given by the Information Sciences Branch of the Office of Naval Research.

CONTENTS

Section		Page
PART I	EXPERIMENTS IN THE USE OF SYNTACTIC INFORMATION IN AUTOMATIC EXTRACTING AND INDEXING	
1	THE SYNTACTIC ANALYZER "PHRASE"	1-1
	1.1 Background: Previous Experiments in the Use of Syntactic Analysis in Automatic Indexing and Extracting	1-1
	1.2 Theory and Methodology	1-4
	1.2.1 Overview	1-4
	1.2.2 Summary of the Four Levels	1-7
	1.2.3 Methods of Ambiguity Resolution at Each Level	1-13
	1.3 Progress and Results	1-29
	1.3.1 Levels 1 and 2 of PHRASE	1-29
	1.3.2 Level 3 of PHRASE	1-35
2	ENGLISH WORD GOVERNMENT	2-1
	2.1 Nature of Word Government	2-1
	2.2 Utilization of Word Government in Syntactic Analysis	2-7
3	DOCUMENTATION	3-1
	3.1 BPHRAS – Level 1 of PHRASE Parser	3-1
	3.2 NESTPH – Level 2 of PHRASE Parser	3-22
	3.3 Output Program – For Level 1 and 2 of PHRASE	3-43
	3.4 CJPHAS Flow Diagram – A Preliminary or Working Draft for Level 3 of PHRASE	3-57
	3.5 Word Government Tables	3-61
4	REFERENCES	4-1
PART II	DESCRIBING AND ABSTRACTING PICTORIAL DATA	
1	INTRODUCTION	1-1
2	REPRESENTATION OF MEANING	2-1
	2.1 Conceptual Classes for Picture Description	2-1

Section			Page
		2.1.1 Case Relations	2-5
		2.1.2 The Remaining Conceptual Classes	2-5
	2.2	Picture Primitives	2-9
		2.2.1 Attribute Primitives	2-9
		2.2.2 Location Primitives	2-10
		2.2.3 Localizing Primitives	2-11
		2.2.4 Qualifying Primitives	2-12
		2.2.5 Classificational Primitives	2-12
		2.2.6 Operative Primitives	2-12
	2.3	Structuring the Data Base	2-13
		2.3.1 The Metanet	2-13
		2.3.2 Relationship Properties	2-13
3	NATURAL LANGUAGE ASPECTS OF CONCEPTUAL MAPPING		3-1
	3.1	Semantic Ambiguity	3-1
	3.2	Parsing Pictorial Descriptions	3-3
4	REFERENCES		4-1

Appendix

A	THE USE OF CASE STRUCTURES IN SEMANTIC MAPPING	A-1
B	DESCRIPTIVE REPRESENTATIONS OF REMOTELY SENSED IMAGE DATA	B-1

ILLUSTRATIONS

Figure		Page
	PART I	
1	Sample of Limited Parsing	1-2
2	Four Levels of Analysis	1-6
3	Relationship of Basic Phrases to Binary Tree Representation	1-9
4	Examples of Participle Usages	1-11
5	Sample of Grammatical Structure After Level 3 Analysis	1-12
6	Resolution of Preposition-Conjunction Ambiguity	1-14
7	Example of Operation of Juxtaposition Rules	1-16
8	Example of All Possible Noun Phrases and Verb Phrases	1-19
9	Portion of Noun-Verb Ambiguity Logic	1-20
10	Example of Noun-Verb Ambiguity Resolution	1-23
11	Examples of Higher Level Phrases	1-26
12	Sample of Parsing Output	1-31
13	Example of Error Correction at Level 4 (Sentence 58)	1-36
14	Sample of Processed Text	1-37
15	BPHRAS Flow Diagram	3-2
16	NESTPH Flow Diagram – Conceptual	3-27
17	NESTPH Flow Diagram – Detail	3-32
18	OUTPUT Skeleton Flow Diagram	3-47
19	OUTPUT Flow Diagram – Detail	3-50
	PART II	
2-1	Example of a Conceptual Net for a Picture Description	2-3
3-1	Description Parsings Using BPHRAS	3-5

TABLES

Table		Page
	PART I	
2-1	Full Government Table for the Verbal Uses of Hand	2-9
	PART II	
1-1	Typical Descriptions for the Various Earth Resources Disciplines	1-2
2-1	Basic Concept Classes of a Description	2-2
2-2	Case Relations Used in Picture Description	2-6
2-3	Expressions for Location	2-7
2-4	Attributes of Objects	2-8
2-5	Properties of Relations Concerned With Locations	2-15
3-1	Entries for Letter "A" From Government Tables Pertinent to Picture Description	3-4
A-1	Cases Used by Fillmore	A-2
A-2	Cases Used by Schank	A-4
A-3	Cases Used by Tesler	A-4