Automatic Abstracting Research at Chemical Abstracts Service[†]

J. J. POLLOCK* and A. ZAMORA

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received July 17, 1975

This paper uses a modified version of the extraction algorithm of Rush, Salvador, and Zamora to describe current research at Chemical Abstracts Service into the automatic generation of abstracts from primary documents. The results suggest that: (a) some subject areas are inherently more amenable than others to automatic extraction; (b) tailoring the algorithm for application to a narrow subject area yields better results than attempting to apply it more generally. The scope and viability of both Rush-Salvador-Zamora algorithm and of automatic extracting in general are also discussed.

INTRODUCTION

Many attempts have been made to abstract original documents by computer but none has succeeded in producing abstracts approaching good manual abstracts in quality. Moreover, given the present state of linguistic theory, it does not seem likely that a program capable of producing abstracts of manual quality will be written in the near future. In spite of this, research into automatic abstracting is still attractive because of the following factors:

- Manual abstracting is expensive and time-consuming.
- Machine-readable journals will probably become more widely available in the near future due to the increasing use of computer-controlled composition in the printing industry. This will provide a relatively cheap and convenient data base for experimentation.
- Given the availability of machine-readable journals, automatic abstracting will be much cheaper and faster than manual abstracting.
- Automatic abstracting produces abstracts already in machine-readable form for later processing steps.
- While automatic abstracts may never be as good as manual abstracts, they may be good enough for practical purposes, especially with simple manual editing.

Research at Chemical Abstracts Service (CAS) on automatic abstracting differs from previous work in several ways. In the past, abstracting programs have usually been applied to very general data bases (e.g., novels, textbooks) or heterogeneous ones (e.g., articles from widely disparate magazines and journals). The original documents used in our work were all abstracted at CAS for *Chemical Abstracts* (CA) (Volume 77, Issues 25 and 26). Many are from the Pharmacodynamics section of CA (Section 1). We consider that this restriction of the data base to a well-defined subject area is important in principle, and that it is unrealistic to expect a single algorithm to be able to abstract documents on a wide range of subjects.

Another difference is that the present work is aimed at producing not just any kind of abstract, but a specific type which will conform to certain CAS standards.

The abstracting program used in this work is a modified version of the Rush-Salvador-Zamora Automatic Document Abstracting Method (ADAM).[1] The algorithm used is essentially the same, but the program has been greatly speeded up and tested on an extensive chemical data base.

ADAM—THE ABSTRACTING PROGRAM

Abstracts and Extracts. An *abstract* may be defined as "an abbreviated, accurate representation of a document" while an *extract* may be said to consist of "one or more portions of a document selected to represent the whole."[2] Using these definitions, we would prefer *automatic extracting* to the traditional term of *automatic abstracting,* since virtually all automatic abstracting research has been confined to selecting sentences from the original document to form an extract. In the case of ADAM, this distinction is a little blurred since, although it does create an extract, editing is performed on the original sentences to produce somewhat different sentences. However, ADAM does not create new sentences, either de novo or by conjoining original sentences, as human abstractors do.

The Algorithm. Most automatic abstracting methods differ from ADAM in two important respects: they rely heavily on statistical criteria as a basis for sentence selection and rejection, and they are designed to *select* sentences for abstracts. In contrast, ADAM uses statistical data only peripherally and is designed for sentence *rejection* rather than selection.

In ADAM, sentence rejection and selection are based mainly on use of cue words (see Word Control List below), relevance of sentence to title, and frequency criteria. The last two are important conceptually as they allow the algorithm to adapt itself, to some extent, to each individual document. Coherence criteria are also used in sentence rejection and selection and in increasing the readability of the abstract.

Characteristics of ADAM Abstracts. ADAM was designed to produce *indicative* abstracts, i.e., abstracts which enable the reader to judge whether or not he needs to read the original document. These abstracts do not substitute for the original document.

ADAM abstracts have the following characteristics:

- Their size is typically 10–20% of that of the original documents (but no arbitrary cutoff is used).
- They use the terminology of the original document.
- They consist of character strings from the body of the text. No equations, footnotes, tables, graphs, figures, etc., are given.
- Preliminary remarks, negative results (unless these are the only results), methodologies of data gathering, explanations, examples, and opinions are excluded.
- Objectives, results, and conclusions are given.

THE WORD CONTROL LIST (WCL)

Format. The Word Control List (WCL) consists of an alphabetically ordered set of words and phrases (collectively referred to as *terms*) with two associated codes, one semantic and the other syntactic. The format of a WCL entry is

[†] Presented before the Division of Chemical Information, 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 8, 1975.

TERM*X*Y

where X is the semantic and Y the syntactic code. Both X and Y are single characters (defined below) and, for any term, one of the two may be undefined (i.e., left blank; see Table I). Multiword terms are found in the text if there are less than four nonterm words between term words. For example, the term *our work shows* would be found in *our brilliant and inventive work clearly shows.* At present, the WCL contains 777 terms.

The Semantic Codes and Sentence Rejection or Selection. The semantic code associated with a term indicates whether or not a sentence containing it is likely to be suitable for inclusion in an abstract. The semantic codes are implemented hierarchically if a sentence contains more than one WCL term [e.g., I overrides A (see Table II)].

Most of the semantic codes in the WCL are negative, i.e., designed to cause rejection of sentences concerned with background information, speculation, etc. For example, terms such as *previous work* and *not important* would have negative codes.

A few terms (e.g., *this study* and *present work*) do have positive codes, indicating that sentences containing them are probably suitable for inclusion in an abstract.

Intersentence reference may also cause rejection or selection of a sentence. If a rejected sentence refers to a previous sentence, this sentence would also be rejected and, similarly, a previously rejected sentence will be restored if an important sentence refers to it. For example, the second sentence below will cause the first one to be rejected.

> Substance X and substance Y form solutions in liquid ammonia. It is well known that these solutions are blue.

Both sentences will be rejected because *known* has a negative semantic code and *these* indicates intersentence reference. (*It* does not indicate intersentence reference as it is closely followed by *that.*)

Table I. Sample of the Word Control List (WCL)

```
NEXT SECTIONS*A
NO ACCURATE*A
NO ATTEMPT*B
NO* *Z
NOR* *Z
NOT ALWAYS*B
NOT BEEN*A
NOT CLEAR*B
NOT IMPORTANT*A
NOT ONLY*F*F
NOT*L*Z
NOTED*A*V
NOTEWORTHY*K
NOW BEEN*K
NOW*B
NOWADAYS*B
NUMEROUS STUDIES*A
O .*F*F
OBSCURE*A
OBVIOUS*A
OBVIOUSLY*A
OF ABOUT*H
OF COURSE*A
OF* *O
OFFER* *V
OFTEN*E
ON THE OTHER HAND*B
ON WHICH*B*P
ON* *P
ONCE*L
ONE CAN*A
ONE OF*B
ONE SUBJECT*A
ONE*E
ONLY*E
OPINION*A
OR*H*
OTHER*E
OTHERS*E
OUR EXPERIMENTS*I
OUR INVESTIGATION*I
OUR OBSERVATIONS*I
OUR RESULTS*I
OUR STUDIES*I
OUR WORK*I
OUR*K*N
OVER* *P
OVERT*B
P .*F*F
PARAGRAPH*A
PARTICULAR*A
PAST*A
PER CENT*A
PERHAPS*A
PERMISSION*M
PERMITTING*B
```

Table II. Semantic Codes Used in WCL

Code	Meaning	Example
M	Supernegative, automatically deletes sentence	Acknowledgment, appreciation
I	Very positive	Our work, reported here
A	Very negative	Previously, obvious
K	Positive	Noteworthy, postulate
B	Negative	However, i.e.
E	Intensifiers or determiners	Many, most, several
L	Introductory quantifiers	A, especially, once
C	Requires antecedent (intersentence reference)	This, these
H	Heads a modifying phrase	What, whose
F	Null term	Abbreviations
G	Assigned by program to indicate intersentence reference or title words	
J	Continuation of previously assigned semantic code	
D	Delete term	

Table III. Syntactic Codes Used in WCL

Code	Description	Code	Description
A	Article	O	OF
C	Conjunction	Q	TO
D	Delete this term	R	AS
F	Null term	V	Verb
J	Continuation of previously assigned code	W	Auxiliary verb
N	Pronoun	X	IS, ARE, WAS, WERE
P	Preposition	Z	Negatives

Words in the title of the original document also influence sentence rejection and selection. Before the body of the document is processed, its title is matched against the WCL and words in the title which are not found in the WCL are assigned a semantic code G. Sentences containing these words are retained in the absence of terms with negative semantic codes.

Syntactic Codes and Coherence. The syntactic codes (see Table III) are used to perform a partial syntactic analysis of each sentence. Theair main use is to classify commas so that contextual inferences may be applied. Commas are divided into four classes: numerical (e.g., in 123,456), clause (used to separate phrases), serial (which follow members of a series), and parenthetical (which delimit dependent clauses). If the second of a pair of commas delimiting a text string is followed by a verb or by *to,* the commas are parenthetical. If a series of one or more commas in a sentence is followed by a comma or conjunction which is not immediately preceded by a preposition or a verb, the commas are serial.

Parenthetical commas cause the text string they delimit to be deleted. Serial commas are masked to prevent confusion with clause commas during later processing but are unmasked for output. Introductory clauses are deleted if they contain a term with semantic code B, H, or L.

The deletion of parenthetical and introductory clauses is important in principle because it means the abstract will not consist of sentences taken verbatim from the original document.

Finally, before a sentence is accepted for the abstract, it is examined via the syntactic codes assigned during WCL-matching to see if it contains a verb; if it does not, it is rejected.

Frequency Criteria. Frequency criteria are employed in a restricted but theoretically important manner. The frequency of occurrence of a term in the original document determines whether the semantic code assigned a priori in the WCL will be accepted or modified. Terms with positive codes are given less positive codes if their occurrence per

Table IV. Examples of Non-Substantive Introductory Phrases

Phrases Beginning With "in"

IN AGREEMENT WITH PREVIOUS WORK
IN ALL OF THESE STUDIES
IN ANY EVENT
IN CONCLUSION
IN CONTRAST
IN SUMMARY
IN OUR EXPERIMENTS
IN OUR INVESTIGATIONS
IN OUR WORK
IN THE PRESENT EXPERIMENT
IN THE PRESENT EXPERIMENTS
IN THE PRESENT PAPER
IN THE PRESENT REPORT
IN THE PRESENT STUDY
IN THE PRESENT WORK
IN THIS EXPERIMENT
IN THIS STUDY

Intersentential Reference Phrases

APPARENTLY
AS A CONSEQUENCE
AS CAN BE SEEN
AS EXPECTED
AS SHOWN IN THE PREVIOUS SECTION
BASED ON THESE DATA
CONCEIVABLY
CONSEQUENTLY
FIRST,
FROM THESE RESULTS
HENCE
HOWEVER
INDEED
LAST,
ON THE OTHER HAND
SECONDLY
THEREFORE
THUS
TO THIS END
UNFORTUNATELY
UNQUESTIONABLY

Phrases Ending With "that"

IT APPEARS THAT
IT APPEARS, THEREFORE, THAT
IT HAS BEEN CONCLUDED THAT
IT HAS BEEN FOUND THAT
IT HAS NOW BEEN FOUND THAT
IT IS APPARENT THAT
IT IS CLEAR THAT
IT IS CONCLUDED BY THE PRESENT INVESTIGATION THAT
IT IS CONCLUDED THAT
IT IS EVIDENT THAT
IT IS FOUND THAT
IT IS INDICATED THAT
IT IS INTERESTING THAT
IT IS POSSIBLE THAT
IT IS SHOWN THAT
IT IS SUGGESTED THAT
IT IS THEREFORE, REASONABLE TO ASSUME THAT
IT MAY BE CONCLUDED THAT
IT MAY THUS BE CONCLUDED THAT
IT MAY BE NOTED THAT
IT SEEMS A REASONABLE CONCLUSION THAT
IT WAS FOUND THAT
IT WAS ALSO FOUND THAT
THE DATA PRESENTED IN THIS REPORT DEMONSTRATE THAT
THE DATA PRESENTED IN THIS REPORT INDICATE THAT
THE DATA REPORTED IN THIS STUDY IMPLY THAT
THE PURPOSE OF THIS REPORT IS TO DEMONSTRATE THAT
THE PURPOSE OF THIS STUDY IS TO DEMONSTRATE THAT
THE PRESENT EXPERIMENTS SHOW THAT
THE PRESENT RESULTS DEMONSTRATE DIRECTLY THAT
THE PRESENT STUDIES HAVE SHOWN THAT
THE PRESENT STUDY HAS DEMONSTRATED THAT
THE PRESENT WORK INDICATES THAT
THE PRESENT WORK SHOWS THAT

thousand words is greater than four, while negative codes are made less negative if it exceeds seven. These criteria tend to decrease positive codes and thus favor smaller abstracts.

The conceptual importance of the frequency criteria is that they adapt the WCL to each individual document. For example, in a document concerned with paper manufacture, the term *this paper* has much less significance than it would in other contexts and its a priori (i.e., manually assigned) semantic code in the WCL would therefore be decreased. Similarly, in a paper on photographic chemistry, the term *negative* has a special meaning and its semantic code would be changed.

Final Editing. As a final step, the abstract is automatically edited, in the ways indicated below, to delete certain nonsubstantive words and phrases which occur at the start of sentences and to abbreviate or replace certain words and phrases according to CAS standards.

Deletion of Nonsubstantive Introductory Words and Phrases from Sentences. Inspection of ADAM extracts revealed that author sentences often consist of an introductory word or phrase followed by a declarative sentence. (The preceding sentence is an example of this, *Inspection . . . that* being the introductory phrase.) Such words and phrases are alien to the style of abstracts and can be removed without loss of information. These nonsubstantive words and phrases divide naturally into three groups: phrases ending in *that,* which are usually followed by a conclusion; phrases beginning with *In,* which often indicate the scope of the associated statement; and words or phrases signifying that the sentence in question is logically connected with a previous sentence. It is interesting to note that many phrases which are effective for sentence selection (e.g., *In our work*) should not appear in the final abstract as they no longer convey any useful information. (See Table IV for examples of the superfluous words and phrases.)

Abbreviation of Words or Phrases. Abbreviations are applied according to "Abbreviations and Symbols used in ACS Publications" (see Table V).

This includes not only the terms actually listed but also many verb, plural, and prefixed variants as well as words ending in certain suffixes (e.g., ological, ographically).

Non-U.S. Spellings. Examples of spellings which differ from the U.S. practice include *sulph* (which is changed to *sulf* in *sulphone, sulphate,* etc.) and *behaviour* (which is replaced by *behavior*).

Replacing Chemical Compounds by Formulas. Replacing compound names with formulas can result in considerable space saving (e.g., substituting NaOAc for sodium acetate saves nine characters), but one must be careful to replace complete compounds. For example, changing nitrobenzene to $PhNO_2$ is good, but replacing trinitrobenzene with $triPhNO_2$ would be highly inappropriate.

THE DATA BASE

A machine-readable data base of 56 papers was created by keypunching original documents. The only essential rules for this keypunching are that each sentence must end in a period followed by two blanks, that blanks must not be inserted into words, and that there must be at least one blank between words.

The commonest subject area was pharmacodynamics, but the papers in other biochemical areas, physical chemistry, organic chemistry, polymer chemistry, inorganic chemistry, and analytical chemistry were also keypunched. The papers varied from short notes to long papers and had between 111 and 3217 words. Some papers were keypunched in full while other papers had sections giving methodological details (e.g., headed "Experimental" or "Methods") and were omitted. The reasons for this practice are described below.

RESULTS

Modification of the Data Base. ADAM is designed to produce indicative abstracts, which should not contain methodological details. Thus, one would not expect sections of papers connected with such details to be fruitful sources of good sentences for abstracts. Accordingly, ADAM was run on two versions of the first batches of papers to be keypunched: complete papers and papers with the sections containing experimental data omitted. The only portions omitted were those clearly headed "Experimental" or "Methods", which would be easily recognizable as such algorithmically in a machine-readable paper produced during computer-controlled composition.

The Experimental sections were found to contribute no useful sentences to the abstracts, and subsequent papers were keypunched without them. This effects considerable savings both in computer time and in keypunching.

Algorithm Performance: Quality of Abstracts. The difficulty of evaluating abstracts in a convenient and theoretically sound manner is discussed below. In our opinion, most of the abstracts produced by ADAM were functionally adequate. They were not as good as abstracts written by professional abstractors, but they contained enough information for the reader to be able to judge whether or not he needed to obtain the original document.

Subject Area, Document Structure, and Abstract Quality. ADAM produces better abstracts from some subject areas than others; the abstracts are affected by the composition of the WCL and the structure of the original documents.

The WCL was originally designed for general English text and later modified to optimize the abstracting of pharmacodynamic papers. Some specialization of vocabulary is necessary for each subject area, and WCL terms appropriate for one subject area may actually be detrimental in another. Thus the current WCL may favor pharmacodynamics at the expense of other topics, and it might be necessary to use different WCL's for each subject area.

ADAM works best with narrative-style documents.

Table V.

ABBREVIATIONS AND SYMBOLS USED IN ACS PUBLICATIONS

A	ampere
Å	angstrom unit
abs.	absolute
abstr.	abstract
Ac	acetyl (CH_3CO, not CH_3COO)
a.c.	alternating current
ACTH	adrenocorticotropin
addn.	addition
addnl.	additional(ly)
ADP	adenosine 5′-diphosphate
alc.	alcohol, alcoholic
aliph.	aliphatic
alk.	alkaline (not alkali)
alky.	alkalinity*
AMP	adenosine 5′-monophosphate
amt.	amount
amu	atomic mass unit
anal.	analysis*, analytical(ly)
anhyd.	anhydrous
AO	atomic orbital
app.	apparatus
approx.	approximate(ly)
approxn.	approximation
aq.	aqueous
arom.	aromatic
assoc.	associate
assocd.	associated
assocg.	associating
assocn.	association
at.	atomic (not atom)
atm	atmosphere (the unit)
atm.	atmosphere, atmospheric
ATP	adenosine 5′-triphosphate
ATPase	adenosinetriphosphatase
av.	average
b.	(followed by a figure denoting temperature) boils at, boiling at (similarly b_{13}, at 13 mm pressure)
bcc.	body centered cubic
BeV or GeV	billion electron volts
BOD	biochemical oxygen demand
μB	Bohr magneton
b.p.	boiling point
Btu	Britislr thermal unit
Bu	butyl (normal)
Bz	benzoyl (C_6H_5CO, not $C_6H_5CH_2$)
c-	centi- (as a prefix, e.g., cm)
cal	calorie
calc.	calculate
calcd.	calculated
calcg.	calculating
calcn.	calculation
CD	circular dichroism
c.d.	current density
CDP	cytidine 5′-diphosphate
chem.	chemical(ly), chemistry
Ci	curie
clin.	clinical(ly)
CM-cellulose	carboxymethyl cellulose
CMP	cytidine 5′-monophosphate
CoA	coenzyme A
COD	chemical oxygen demand
coeff.	coefficient
com.	commercial(ly)
compd.	compound
compn.	composition
conc.	concentrate
concd.	concentrated
concg.	concentrating
concn.	concentration
cond.	conductivity*
const.	constant
contg.	containing
cor.	corrected
CP	chemically pure
crit.	critical
cryst.	crystalline (not crystallize)
crystd.	crystallized
crystg.	crystallizing
crystn.	crystallization
CTP	cytidine 5′-triphosphate
d-	deci- (as a prefix, e.g., dl)
d.	density (d^{13}, density at 13° referred to water at 4°; d^{20}_{20}, at 20° referred to water at the same temperature)
D	debye unit
d.c.	direct current
DEAE-cellulose	diethylaminoethyl cellulose
decomp.	decompose
decompd.	decomposed
decompg.	decomposing
decompn.	decomposition
degrdn.	degradation
deriv.	derivative
det.	determine
detd.	determined
detg.	determining
detn.	determination
diam.	diameter
dil.	dilute
dild.	diluted
dilg.	diluting
diln.	dilution
dissoc.	dissociate
dissocd.	dissociated
dissocg.	dissociating
dissocn.	dissociation
distd.	distilled
distg.	distilling
distn.	distillation
DMF	dimethylformamide
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
d.p.	degree of polymerization
dpm	disintegrations per minute
DPN	diphosphopyridine nucleotide (NAD)
DPNH	reduced DPN
DTA	differential thermal analysis
ED	effective dose
elec.	electric, electrical(ly)
emf.	electromotive force
emu	electromagnetic unit
en	ethylenediamine (used in Werner complexes only)
EPR	electron paramagnetic resonance
equil.	equilibrium(s)
equiv	equivalent (the unit)
equiv.	equivalent
esp.	especially
ESR	electron spin resonance
est.	estimate
estd.	estimated
estg.	estimating
estn.	estimation
esu	electrostatic unit
Et	ethyl
eV	electron volt
evap.	evaporate
evapd.	evaporated
evapg.	evaporating
evapn.	evaporation
examd.	examined
examg.	examining
examn.	examination
expt.	experiment
exptl.	experimental(ly)
ext.	extract
extd.	extracted
extg.	extracting
extn.	extraction
F	farad
FAD	flavine adenine dinucleotide
fermn.	fermentation
fcc.	face centered cubic
FMN	flavine mononucleotide
f.p.	freezing point
FSH	follicle-stimulating hormone
G	gauss
G-	giga-(10^9)
g	gram
(g)	gas, only as in $H_2O(g)$
g	gravitation constant
GDP	guanosine 5′-diphosphate
GMP	guanosine 5′-monophosphate
GTP	guanosine 5′-triphosphate
H	henry
ha	hectare
Hb	hemoglobin
hr	hour
Hz	hertz (cycles/sec)
ICSH	interstitial cell-stimulating hormone
ID	infective dose
IDP	inosine 5′-diphosphate
i.m.	intramuscular(ly)
IMP	inosine 5′-monophosphate
inorg.	inorganic
insol.	insoluble
i.p.	intraperitoneal(ly)
ir	infrared
irradn.	irradiation
ITP	inosine 5′-triphosphate
IU	International Unit
i.v.	intravenous(ly)
J	joule
k-	kilo- (as a prefix, e.g., kg)
l.	liter
(l)	liquid, only as in $NH_3(l)$
lab.	laboratory
LCAO	linear combination of atomic orbitals
LD	lethal dose
LH	luteinizing hormone
liq.	liquid
lm	lumen
lx	lux
m-	milli- (as a prefix, e.g., mm)
m	meter
m.	melts at, melting at
m	molal
M-	mega- (10^6)
M	molar
manuf.	manufacture
manufd.	manufactured
manufg.	manufacturing
math.	mathematical(ly)
max.	maximum(s)
Me	methyl (not metal)
mech.	mechanical(ly)
metab.	metabolism
min	minute (time)
min.	minimum(s)
misc.	miscellaneous
mixt.	mixture
MO	molecular orbital
mol.	molecule, molecular (not mole)
m.p.	melting point
μ	micron; also micro- (as a prefix, e.g., μl)
MSH	melanocyte-stimulating hormone
Mx	maxwell
n-	nano- (10^{-9})
n	refractive index (n^{20}_{D} for 20° and sodium D light)
N	newton
N	normal (as applied to concn.)
NAD	nicotinamide adenine dinucleotide (DPN)
NADH	reduced NAD
NADP	nicotinamide adenine dinucleotide phosphate (TPN)
NADPH	reduced NADP
neg.	negative(ly)
NMN	nicotinamide mononucleotide
NMR	nuclear magnetic resonance
no.	number
NQR	nuclear quadrupole resonance
obsd.	observed
Oe	oersted
Ω	ohm
ORD	optical rotatory dispersion
org.	organic
oxidn.	oxidation
P	poise
p-	pico- (10^{-12})
p.d.	potential difference
Ph	phenyl
phys.	physical(ly)
PMR	proton magnetic resonance
polymd.	polymerized
polymg.	polymerizing
polymn.	polymerization
pos.	positive(ly)
powd.	powdered
ppb	parts per billion
ppm	parts per million
ppt.	precipitate
pptd.	precipitated
pptg.	precipitating
pptn.	precipitation
Pr	propyl (normal)
prep.	prepare
prepd.	prepared
prepg.	preparing
prepn.	preparation
prodn.	production
psi	pounds per square inch
psia	pounds per square inch absolute
psig	pounds per square inch gage
purifn.	purification
py	pyridine (used in Werner complexes only)
qual.	qualitative(ly)
quant.	quantitative(ly)
R	roentgen
redn.	reduction
ref.	reference
rem	roentgen equivalent man
rep	roentgen equivalent physical
resp.	respective(ly)
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
RQ	respiratory quotient
(s)	solid, only as in AgCl(s)
sapon.	saponification
sapond.	saponified
sapong.	saponifying
sat.	saturate
satd.	saturated
satg.	saturating
satn.	saturation
s.c.	subcutaneous(ly)
SCE	saturated calomel electrode
SCF	self-consistent field
sec	second (time unit only)
sec	secondary (with alkyl groups only)
sep.	separate(ly)
sepd.	separated
sepg.	separating
sepn.	separation
sol.	soluble
soln.	solution
soly.	solubility*
sp.	specific (used only to qualify physical constant)
sp. gr.	specific gravity
sr	steradian
St	stokes
std.	standard
sym.	symmetrical(ly)
T-	tera- (10^{12})
TEAE-cellulose	triethylaminoethyl cellulose
tech.	technical(ly)
temp.	temperature
tert	tertiary (with alkyl groups only)
theor.	theoretical(ly)
thermodn.	thermodynamic(s)
THF	tetrahydrofuran
titrn.	titration
TPN	triphosphopyridine nucleotide (NADP)
TPNH	reduced TPN
Tris	tris(hydroxymethyl)aminomethane
TSH	thyroid-stimulating hormone
UDP	uridine 5′-diphosphate
UMP	uridine 5′-monophosphate
USP	United States Pharmacopeia
UTP	uridine 5′-triphosphate
uv	ultraviolet
V	volt
vol.	volume (not volatile)
W	watt
wt.	weight

Plurals of noun abbreviations are formed by adding "s" to the singular abbreviation except when a single abbreviation is designated to show both the singular and plural forms and except for words marked * whose plurals are not abbreviated. Verb forms that require "s" are treated similarly. Words formed by adding prefixes to words normally abbreviated are also abbreviated, as microchem. for microchemical. Other well established abbreviations, as etc., i.e., e.g., and abbreviations for English units of weight and measure, are also used. Unit abbreviations signify both singular and plural forms. Words ending in -ology or -ological(ly) are abbreviated -ol., e.g., geol. for geology. Words ending in -ography or -ographic(al)(ly) are abbreviated -og., e.g. chromatog. for chromatographic.

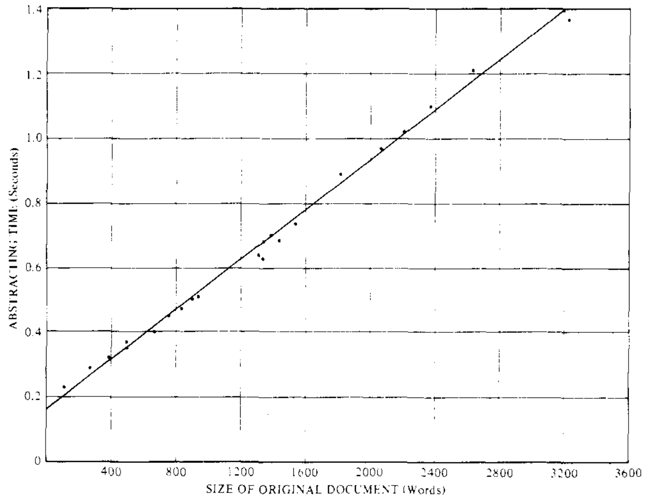

Figure 1. Program performance.

Many pharmacodynamic papers are written in this way; they begin with an introduction giving background information and stating the purpose of the work, follow this with "Methods" and "Results" sections, and end with a series of conclusions. This "linear" structure is well suited to the production of a coherent abstract as sentences are almost inevitably extracted in a logical narrative sequence. Other subject areas may be less linearly reported. For example, a physical chemistry paper may consist of a few measurements followed by a series of discussions, an organic chemistry paper may describe a number of syntheses, or an analytical paper may detail a recommended procedure. In these cases, there may be no sentences which adequately summarize the whole work. A human abstractor can summarize a discussion with a single sentence, but it is not feasible for a program to do this.

Also, some subject areas are more dependent on nontextual information than others. For example, the text of an organic chemistry paper may be incomprehensible without the accompanying structures and a physical chemistry document may be replete with complex equations without which the text makes little sense.

Program Speed. During our investigation of the execution speed of ADAM, we found that the time required to abstract each original document was proportional to the square (N^2) of the number (N) of words it contained. The first step of ADAM, in which the words of the original document were sorted alphabetically, initially used a sorting process in which the execution time was proportional to N^2. Since this initial sort was by far the most time-consuming part of the abstracting process, the initial sorting process was changed to one in which the execution time was dependent on $N \log N$. This caused a dramatic decrease in abstracting time; for example, the time required for the longest paper (3217 words) decreased from 23 to 1.4 sec, while the average amount of IBM 370/168 computer time per paper fell from 4.38 to 0.59 sec. The contribution of the sorting process to the total abstracting time now appears to be much smaller. This assumption is supported by Figure 1 which shows that the abstracting time required by a document is approximately proportional to the number of words in it. This virtually linear dependence on N makes the program attractive economically since processing large papers is no longer disproportionately expensive.

Abstract Size. On the average, ADAM abstracts contained 19% as many words as the original documents. This is longer than we would wish but still a very substantial reduction in size. In many cases, overlong abstracts could be significantly reduced in size by the removal of entire, inappropriate sentences. This would be a very simple manual editing task, not demanding intimate knowledge of the subject area.

The Problem of Abstract Evaluation. At first, this problem seems almost trivial. If it is possible to produce abstracts by computer, then surely it must be possible to determine how good they are? Moreover, without a reliable method of judging the quality of the abstracts, how can one modify the algorithm to produce better output? Here we would briefly like to review previous work on abstract evaluation and offer our own conclusions.

Methods for evaluating abstracts can be classified as intuitive, statistical, computational, and functional.

The *intuitive* method consists simply of human judgment of the abstract. It is the most popular method because of its simplicity, and it is widely used by automatic abstracting researchers and in training abstractors. It has the disadvantages of being inconsistent, time-consuming, and nonquantitative.

One important *statistical* method involves creating "ideal" extracts (composed of sentences chosen by professional abstractors) and correlating these statistically with the corresponding automatic extracts. The problem with this method lies in the assumption that there is a single ideal extract for each paper. In fact, there may be several good extracts for a given paper or none. It has also been shown experimentally[3-5] that each abstractor produces a different extract for a given paper, and that if an abstractor is given the same paper after a lapse of eight weeks, he will produce a substantially different abstract. In short, this method also relies heavily on human judgment and shares the defects of the intuitive method. It is also possible to statistically compare the vocabulary of the original document to that of the abstract to determine how representative the latter is. However, this tells little about how good the abstract is.

A *computational* method of evaluating abstracts would be extremely useful as it would almost certainly be very much faster and more convenient than manual methods. We investigated the possibility of evaluating abstracts by comparing the amounts of information in the original document and the abstract but were forced to conclude that this is not feasible. The main reason for this is that not enough is known about linguistics or the nature of information. For example, what is meant by the *information content* of a document? Does this mean only that information explicitly stated or does it include statements which follow logically from these? Is the amount of information in a document constant? It could be argued that it varied with the knowledge and intelligence of the reader. It could also be argued that an identical text passage written by different authors would have a different meaning in each case, varying with the assumptions of the writer. In order to evaluate abstracts meaningfully, it would be necessary to distinguish algorithmically between different kinds of information (new, old, trivial, important, etc.) and different levels of information (haloalkyl aryl ketones vs. bromodifluoropentyl naphthyl ketones). Should the above difficulties be overcome, then the problem of quantification arises. In what units should information be expressed and how would the numerical value of each piece be computed? Would the quantitative values depend on the type of information or the reader or author? It is not inconceivable that an algorithmic solution to these problems will eventually be developed, but we do not think it likely in the near future.

Various *functional* methods for evaluating abstracts have been tried in the past. Payne, Altman, and Munger[6,7] asked two groups of students to answer questions about a document after having read the document itself or its abstract. Resnick, Rath, and Savage[8,9] used a similar method to evaluate the usefulness of two types of abstracts compared to that of the complete text or just the title of documents. Abstracts have also been evaluated on the basis of index term content[10] and by their retrieval capabilities.[11]

For indicative abstracts, actual users of abstracts could be asked to read abstracts, decide if they needed the original document, read the original document, and review their decision. This would indicate whether or not the abstracts fulfilled their function and feedback might help to improve the abstracting process. Such a procedure would, of course, be slow and inconsistent.

Possibly the difficulty of automatic evaluation of abstracts is rooted in the same problems as automatic abstracting, and it may not be feasible to solve the one without the other.

CONCLUSIONS

- The quality of ADAM abstracts, while lower than that of good manual abstracts, is functionally adequate.
- ADAM requires, on the average, 0.6 sec of IBM 370/168 computer time per document.
- Since it is likely that more and more journals will become available in machine-readable form in the future, automatic abstracting is desirable because it is potentially much faster and cheaper than manual abstracting.
- Automatic extracting algorithms are most suited to documents with a linear, narrative structure and therefore are more successful with some subject areas than others.
- ADAM needs a specialized WCL for each subject area.
- ADAM abstracts can be improved by simple manual editing.
- It will not be possible to produce abstracts of manual quality by computer without a great breakthrough in linguistics, especially in the area of semantics.
- No theoretical basis now exists by which an abstract can be quantitatively evaluated by comparing its information content with that of the original document.

FUTURE WORK

Future work will include improving ADAM, possibly by modifying the algorithm to give greater weight to some sections of the original document than others (introducing location criteria), by applying it to a machine-readable journal, and by studying areas other than pharmacodynamics in depth to see how much specialization of the WCL is needed.

At present, the frequency thresholds at which the a priori semantic codes in the WCL are changed are the same for each term and are based on intuition. A statistical study of the vocabulary of each subject area will provide an individual frequency threshold for each term which will reflect more precisely its actual occurrence in text.

APPENDIX. ADAM ABSTRACTS

In the following examples, sentences shown in italics in the abstracts are those which we would expect a human editor to remove *in toto*; those which we consider essential are in boldface. In each example, the "experimental" section of the original document was not abstracted.

Structure-Toxicity Relationships of Substituted Phenothiazines.

Charles H. Nightingale, Melissa Tse, and Elliot I. Stupak CA-77-25-160007A

Phenothiazine hydrophobicity is related to pharmacol. response, i.e., the greater the hydrophobicity the greater the activity. Relatively complex pharmacol. effects were monitored but no attempt was made to elucidate the role of the absorptive process in modifying pharmacol. response. A relatively simple and inexpensive biol. test system can be utilized to correlate pharmacol. effect with the phys.-chem. properties of phenothiazines and to study the effect of the absorptive process in modifying such response. As stated by Zografi and Munshi, the phenothiazine mol. appears to be oriented at the interface with the ring toward the nonpolar phase and the alkylamino group directed toward the bulk aq. phase. Changes in hydrophobicity of the ring structure will, therefore, change surface activity significantly. Time of death detns. indicate that the greater the surface activity or partition coeff., the greater the toxicity. *A relationship exists between phenothiazine absorption and hydrophobicity as indicated by partitioning into dodecane.* Although a rank order correlation was also found between surface pressure and time of death for the 2-substituted derivs., a similar relationship could not be demonstrated for the 1-, 2-, and 3-chloro analogs. The goldfish test system is capable of discerning structure-toxicity relationships cf substituted phenothiazines. Differences in the appearance of phenothiazine-induced toxicity depend upon the ability of the free drug to partition into the fish and this process can be correlated with dodecane partition coeffs. *One must use caution in extrapolating these results to tranquilizing activity in higher animals since inherent activity differences between the various phenothiazines were reported.*

COMMENT: The abstract is too long (23% of the original document) because too much background and trivial information is given. However, the main point of the paper (the use of the goldfish test system) is well emphasized (by being mentioned repeatedly) and the abstract faithfully reflects the content of the paper.

STATISTICS: Full original document = 1138 words.
ADAM abstract = 263 words.
Boldface abstract = 215 words.
CA abstract = 70 words.

Pseudoephedrine and the Dog's Eustachian Tube.
J. Edward Dempsey and Richard To. Jackson CA-77-25-160116K.

This study is an attempt to give an objective evaluation of the effectiveness on the dog's eustachian tube of two widely-used otolaryngic drugs-pseudoephedrine hydrochloride (sudafed) and triprolidine hydrochloride (actidil). *Merck, in 1893, obtained an isomeric alkaloid from the European E. vulgaris which he called pseudoephedrine. The pharmacol. properties of ephedrine and its isomers were largely established in the 1920's. Surprisingly, there are only a few studies of the effects of the ephedrine isomers on otolaryngic tissues-none on the eustachian tube. The four ephedrine isomers are d(-) ephedrine, l(+) ephedrine, d(-) pseudoephedrine and l(+) pseudoephedrine. Triprolidine is an antihistamine.* Two kinds of dose-response curves were obsd. We were unable to evoke a response with a single i.v. injection. *Our results are not as clear if the drug is given i.v.* In every instance tested in our expts., after complete tachyphylaxis to pseudoephedrine, repeated arterial administration 0.2 mg. of tyramine produced a normal series of responses. We found a drug, which is used daily, exhibiting tachyphylaxis. *We have seized on a particular datum as an indication.*

COMMENTS: The abstract contains a great deal of irrelevant information. Sentences which contain dates from 1959 onwards are usually rejected by ADAM but 1893 and 1920 are not in the WCL. Sentences like *Tripolidine is an antihistamine* illustrate the difficulty of distinguishing between different kinds of information. The statement is correct and contains a good deal of information per character but it is not suitable for an abstract because it is well-known to workers in the field. Sentences such as *Our results are not as clear if the drug is given i.v.* show the fallibility of relying on the surface patterns of language. The phrase *our results,* which usually occurs in important statements, is here attached to a trivial one.

STATISTICS: Full original document = 1596 words.
ADAM abstract = 174 words.
Boldface abstract = 90 words.
CA abstract = 54 words.

Plasma Levels and Absorption of Methaqualone after Oral Administration to Man.

Robert N. Morris, Gwendolyn A. Gunderson, Steven W. Babcock, and John F. Zaroslinski CA-77-25-160000T.

The present study was designed to det. human blood levels of methaqualone after oral administration of a com. prepn. of the drug, and to est. the rate and extent of absorption. The plasma elimination half-life detd. from the graph was 2.6 hr. The corresponding elimination rate const. was 0.267 hr. *The ratios obtained at each time point up to 3 hr are shown in the sixth column of table II.* 67% of the dose was absorbed within one hr and 99% in 2 hr. *Insufficient data for the absorptive phase of the curve presented in fig 2 are available to accurately est. the over-all rate const. for dissoln. and absorption.* The corresponding absorption half-time is 0.6 hr. *The value, K_a in this case represents the combined rate for dissoln. and absorption, insofar as the drug was administered orally in a solid dosage form. One subject was heavily sedated but did not sleep. The peak blood levels of these 3 were among the highest in the group.* The onset of the signs occurred between 15 and 40 min after drug. All subjects appeared alert and reported no symptoms of sedation after about 4 hr. Plasma levels indicated that 80 to 90% of the drug is cleared from the plasma within 8 hr. The plasma half-life of 2.6 hr agrees well with that estd. from the data reported by Berry. *Wide variations in plasma levels between individuals such as those noted in the present study were also reported by Berry, but he did not described the conditions in which the drug was given and reported plasma levels in only 3 patients receiving methaqualone alone.*

COMMENTS: This contains more than enough information for an indicative abstract but much of the data chosen is the same as that chosen by the CA abstractor. References are made to absent graphic data and some trivial information is given.

STATISTICS: Full original document = 1606 words.
ADAM abstract = 272 words.
Boldface abstract = 133 words.
CA abstract = 77 words.

Spectrophotometric Determination of Dimethyl Sulfoxide
Z. Dizdar, Z. Idjakovic. CA-77-26-172392K.

Working with aq. solns. of dimethyl sulfoxide(DMSO) has imposed the need for a quick and simple method for its quant. detn. *The method of potentiometric titrn. used for the detn. of the sulfoxide dissolved in Ac_2O, when the titrn. is performed with an Ac_2O or dioxan soln. of $HClO_4$, cannot be applied to such solns. either. The oxidn. methods fail in the presence of reducing agents, e.g., Me_2S which is often present in the sulfoxide.* As a result, a change should be expected in the position of the absorption spectra, which are shifted to longer wavelengths. *This medium effect has long been known.* Though the shapes of the spectra are not essentially modified by the change of the medium, the intensities of the bands are increased and red shifts obsd. in the presence of DMSO. As the best result was obtained with ammonium iron(III) sulfate the spectrophotometric detn. of mg quantities of DMSO in aq. solns. was developed with this salt as the reagent. DMSO is quant. recovered from the resin, which is in agreement with an earlier finding, and also proves the correctness of the method used for elimination of interfering cations. Sensitivity: the molar absorptivity is 3.1 mole-1 at 419 nm. The min. amt. detectable is 260 mug/cm^2.

COMMENTS: The abstract contains a great deal of trivial and background information. The useful information given is essentially the same as that chosen by the professional abstractor for CA.

STATISTICS: Full original document = 884 words.
ADAM abstract = 220 words.
Boldface abstract = 149 words.
CA abstract = 114 words.

Antispasmodics Derived from Aminopyrimidines.

P. K. Jesthi and M. K. Rout CA-77-25-160005Y.

The following types of compds. derived from various amino-pyrimidines have been prepd.: betadiethylacetamido, betaamino ethylamino, beta-diethylamino ethylamino-, morpholino ethylamino , (5) piperidino ethylamino-. For prepg. the compds. of type (1) the corresponding amino-pyrimidines were condensed with $ClCH_2COCl$ and the resulting products were made to condense with Et_2NH under appropriate conditions. Altogether ten representative members of the different type of compds. prepd., in the present investigation have been screened for their antispasmodic and antihistaminic properties. The tests were performed on strips of guinea-pig ileum. Compd. no. 3 of table V was the most active antispasmodic and inhibited 50% of the spasm produced by std. dose acetylcholine in a dose of 162 mug./ml. Compd. no. 1 of table II was found to be most active anthihistaminic and inhibited 50% of the spasm produced by std. dose of histamine acid phosphate in a dose of 200 mug./ml.

COMMENTS: This abstract illustrates a characteristic difficulty in extracting synthetic organic text: the indispensability of the graphic material of the paper. For example, the reader cannot deduce the structure of compound no. 3 of Table V. In spite of this, the reader should receive an accurate impression of the paper and be able to decide whether or not he wants to read it.

STATISTICS: Full original document = 884 words.
ADAM abstract = 143 words.
CA abstract = 76 words.

Effects of Vasoactive Agents and Diuretics on Isolated Superfused Interlobar Renal Arteries. J. W. Strandhoy, R. Cronnelly, J. P. Long and H. E. Williamson. CA-77-25-160038M.

Little information is available concerning the actions of diuretics and other drugs on renal vascular smooth muscle, uncomplicated by reflex, hormonal or other extraneous influences. The purpose of this study therefore was to isolate a segment of the renal vascular tree for evaluation in vitro of direct actions of diuretics and other vasoactive agents. Two diuretics were evaluated for activity on the interlobar artery. No relaxant effects were obsd. with either agent. No change in base line tension was obsd. The isolated intralobar artery was found to contract in response to KCl and symmpathetic nerve stimulation. No evidence was found to indicate that tension of larger renal arteries is decreased in response to furosemide. Superfusion of the tissue with hydrochlorothiazide was found to attenuate the contractile responses of norepinephrine. The highest doses of norepinephrine used were not antagonized sufficiently suggesting a competitive depression of reactivity. Hydrochlorothiazide was found to produce dose-related contractions of the interlobar strips. High doses were necessary. The interlobar arteries contribute to the renal vasoconstriction produce by norepinephrine and dopamine, but do not contribute to the renal vasodilation produced by furosemide or small doses of dopamine.

COMMENTS: Too much background, negative, and trivial information is given but the abstract still contains enough information for an indicative abstract.

STATISTICS: Full original document = 1502 words.
ADAM abstract = 199 words.
Boldface abstract = 165 words.
CA abstract = 83 words.

LITERATURE CITED

(1) Rush, J. E., Salvador, R., and Zamora, A., "Automatic Abstracting and Indexing. II. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria," *J. Am. Soc. Inf. Sci.*, **22** (4), 260–74 (1971).

(2) Weil, B. H., "Standards for Writing Abstracts," *J. Am. Soc. Inf. Sci.*, **22** (4), 351–7 (1970).

(3) Rath, G. J., Resnick, A., and Savage, T. R., "The Formation of Abstracts by the Selection of Sentences. Part I. Sentence Selection by Men and Machines," *Am. Doc.*, **12** (2), 139–41 (1961).

(4) Resnick, A., "The Formation of Sentences by the Selection of Sentences. Part II. The Reliability of People in Selecting Sentences," *Am. Doc.*, **12** (2), 141–3 (1961).

(5) Resnick, A., and Savage, T. R., "The Consistency of Human Judgments of Relevance," *Am. Doc.*, **15** (2), 93–5 (1964).

(6) Payne, D., Altman, J., and Munger, S. J., "A Textual Abstracting Technique, Preliminary Development and Evaluation for Automatic Abstracting Evaluation," American Institute for Research, Pittsburgh, Pa., 1962 (AD 285 032).

(7) Payne, D., "Automatic Abstracting Evaluation Support," American Institute for Research, Pittsburgh, Pa., 1964 (AD 431 910U).

(8) Rath, G. J., Resnick, A., and Savage, T. R., "Comparisons of Four Types of Lexical Indicators of Content," *Am. Doc.*, **12** (2), 126–30 (1961).

(9) Resnick, A., "Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents," *Science*, **134** (3438), 1004–6 (1961).

(10) Caras, G. J., "Indexing from Abstracts of Documents," *J. Chem. Doc.*, **8** (1), 20–22 (1968).

(11) "Final Report on the Study for Automatic Abstracting," Thompson Ramo Wooldridge, Inc., Canoga Park, Calif., 1961 (PB 166 532).

Computer Programs for Editing and Validation of Chemical Names[†]

GERALD G. VANDER STOUW

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received July 31, 1975

Chemical Abstracts Service (CAS) has developed computer procedures for editing chemical names, including both CA Index Names and names from the original literature. These editing programs include steps which will automatically correct errors in punctuation, format, capitalization, and italicization where possible, as well as error detection steps which generate diagnostic messages. Nomenclature translation processes have been incorporated into these editing programs in order to detect errors in names generated as CA Index entries and to validate such names by comparing them to their structural records.

Chemical Abstracts Service (CAS) has long paid considerable attention to careful editing of material which is going into its publications, and has attempted to correct errors before they are published and thus avoid scattering or loss of information. As the computer-based publication system of CAS has developed during the last several years and as the volume of material being processed has grown rapidly, considerable emphasis has been placed on using computer programs to reduce the amount of human effort spent in editing data that enter the system. Some of these edits have previously been described.[1-3] The development of these edits has emphasized that data which are correct should be verified as early as possible in the system, so that data which have been keyboarded and edited can be placed on file in a correct form and brought out only as needed for inclusion in particular products. Our goal has been to develop error detecting edits which are sufficiently effective so that items which pass these edits do not require human review; only the questioned data (i.e., those items which do not pass the edits) need to be reviewed.

CHEMICAL SUBSTANCE PROCESSING

The steps involved in chemical substance processing in the CAS publication system have recently been described.[4,5] In the CAS system (see Figure 1), there are basically two places where chemical names are recorded in computer-readable form. First, if a substance selected as a Chemical Substance Index entry has a name provided in the source document, Name Match can be attempted; that is, the name given in the document can be matched against substance names already on file. If a match occurs, the CAS Registry Number and CA Index Name are retrieved. Names input for Name Match represent a relatively uncontrolled vocabulary. They are not formulated according to the rigid rules used for CA Index Names and therefore cannot be subjected to as rigorous editing, but some computer edits are applied to them.

If a substance cannot be identified by Name Match, a

[†] Presented to the 168th National Meeting of the American Chemical Society, Atlantic City, N.J., Sept 11, 1974.