5, respectively, adds nothing at all to the precision of this search.

On the other hand, molecules composed entirely of common chemical groups could not be defined precisely by the origial version of the Fragmentation Code. For searches involving simple aliphatic and carbocyclic molecules, the effort spent in encoding the molecule in detail is well repaid by the increase in precision the new descriptors provide. The molecule in Figure 8 is amphetamine. A search in the same portion of the file with the 1963 version of the Fragmentation Code in statement 2 yields 987 patent families. The new descriptors introduced in 1970 reduce the number of postings in statement 3 to 162. The 1972 version of the Code gives only 53 postings in statement 4, and the 1981 version gives only 39 in statement 5, which is comparable with the number of patents we retrieved for cimetidine.

This improved precision reduces the number of false drops only in the later parts of the file, of course. A search for all references to amphetamine in the whole of files WPI and WPIL would not be at all precise. Most simple molecules such as amphetamine have been known for a long time, and searches that involve such molecules usually involve some other inventive feature. If the other feature of the invention is a second compound, the search can be very precise indeed. The cimetidine search in Figure 7 was saved and recalled in statement 6, and the two searches were combined with the AND operator in statement 7. The only patent this combined strategy retrieved is directed to a composition comprising precisely the two compounds we were looking for. Had we combined the cimetidine search with statement 2, we would have obtained only one more posting, which is quite an improvement over the 987 documents retrieved by the strategy based upon amphetamine alone!

The Fragmentation Code can be used for searching the CPI files whenever molecular structure is an important feature of the invention being searched. For searches involving chemical processes, simple aliphatic and carbocyclic compounds, or nonpharmaceutical, nonagricultural compositions containing well-known compounds, it is usually not the most efficient way to search. But for complex molecules, especially heterocyclic and organometallic compounds, nothing else retrieves so many relevant references and so few false drops and does it so quickly. If you have not been using it, you probably do not know how many references you have been missing.

REFERENCES AND NOTES

(1) Ex parte Markush, 340 *OG* 839, 1925 *CD* 125.
(2) Norton, P. "Central Patents Index (CPI) as a Source of Information for the Pharmaceutical Chemist". *Drug Inf. J.* **1982**, 208-215.

Semiautomatic Indexing of Structured Information of Text

FUJIO NISHIDA,* SHINOBU TAKAMATSU, and YONEHARU FUJITA
Department of Electrical Engineering, Faculty of Engineering, University of Osaka Prefecture, Mozu-Umemachi, Sakai, Osaka, Japan 591

Received June 3, 1983

This paper presents a method of semiautomatic information extraction from text such as patent claim sentences or summaries of technical papers written in English as well as in Japanese. The input sentences are parsed, reduced, and normalized into almost the same form of the internal expressions for both the languages except terms. Subsequently, specified information is extracted in a specified language of English or Japanese.

INTRODUCTION

Automatic text processing has been studied actively with the advance of computer technology. The first main linguistic work was the automatic indexing of text, and various approaches of text analysis were established on the basis of statistical distribution of characteristic and technical terms involved in text.[1,2]

Subsequently, structural analysis of text has been introduced.[5-11] It is based on the concept of case and frame presented by Fillmore and Minsky[3,4] and aimes at processing of structured information or knowledge involved in text. Extacting, storing, and handling of structured information will be essential to knowledge engineering and information science in the near future.

This paper describes the outline of a method of semiautomatic information extraction from texts such as patent claim sentences and summaries of technical papers written in English as well as in Japanese. The information to be extracted is designated by a specification table as shown under Specification for Extracted Information. From the practical viewpoint, it is desired but difficult to give a general specification that designates special knowledge to be extracted from texts of various special fields. In this paper, the specification table is given, for simplicity, for each specific field, and each kind of the main subject of technical papers though the scheme of the specification tables is almost the same.

Each sentence of the text is parsed semiautomatically with precise syntactic and general semantic information. If certain serious dependency ambiguities of terms are found that cannot be resolved by the ordinary category reference, the specific knowledge of the subframe associated with the term is retrieved from a knowledge database to resolve them. If the ambiguity cannot be resolved yet, then the system asks the user the correct dependency relation among terms.

After the parsing, the internal expression is normalized into a form similar to the specification. The specified information is extracted by scanning the normalized internal expressions of the text several times and stored in a form of relational tables or modified inverted files.

The case labels and the category names appearing in the internal expressions of technical sentences as well as in the subframes of specification tables were shared in both English and Japanese. Thereby, the extracted information can be immediately represented by the terms of a specified language by means of term-by-term replacement without serious deviation of meaning.

SPECIFICATION FOR EXTRACTED INFORMATION

The information to be extracted is designated by a specification table. It consists of several subframes related to the subject term:

$$L: \ (K_1\text{-}C_1:\_, \ldots, K_i\text{-}C_i:\underline{t}, \ldots, K_n\text{-}C_n:\_) \qquad (1)$$

0095-2338/84/1624-0015$01.50/0

Table I. Specification Tables

(A) For Semiconductor Devices

```
FUNCTION: (PRED-PhysicalACTion:_ , AGent-PRODUCTS: t,
              OBJect-PHYSOBJ ∪ PHYSQUANT:_,
           CONDition:_ )
PROPERTY: (PRED-ATTRibute ∪ RELation:_, OBJ-PHYSQUANT
          ∪ PROPerty:_, LOCation-PRODUCTS: t,
          DEGRee-VALue:_, PARTICipant-PHYSQUANT ∪
                PROP:_ (OBJ-PHYSQUANT ∪ PROP: *,
                        POSSESSOR-PHYSOBJ:_ ))
COMPOSITION: (OBJ-PRODUCTS: t, COMPonent-PRODUCTS:_ )
LOCATION: (OBJ-PRODUCTS: t, LOC-PRODUCTS:_ )
PROCESS: (PRED-PACT:_ , OBJ-PHYSOBJ:_ , GOal-PRODUCTS: t,
             INSTRument-PHYSOBJ:_,
             LOC-PHYSOBJ ∪ PHYSLOC:_, COND:_,
          MEANS-PACT:_ )
```

(B) For the Manufacturing Processes of Semiconductor Devices

```
PROCESS: (PRED-PACT: t, OBJ-PHYSOBJ:_, GO-PRODUCTS:_,
          INSTR-PHYSOBJ:_, LOC-PHYSOBJ ∪ PHYSLOC:_,
             COND:_ )
PRINCIPLE: (PRED-PACT: t, MEANS-PRINCIPLE ∪ PACT:_)
COMPOSITION: (OBJ-PACT: t, COMP-PACT:_  )
```

(C) For the Properties of Semiconductor Devices

```
ANALYSIS; (PRED-THINKACT:_, OBJ-PHYSQUANT ∪ PROP:t
                            (OBJ-PHYSQUANT ∪ PROP: *,
                             POSSESSOR-PRODUCTS:_ ),
           COND:_, MEANS-PRINCIPLE ∪ THINKACT:_ )
PROPERTY; (PRED-ATTR ∪ REL:_, OBJ-PHYSQUANT ∪ PROP:t,
           LOC-PRODUCTS:_, DEGR-VAL:_,
           PARTIC-PHYSQUANT ∪ PROP:_
           (OBJ-PHYSQUANT ∪ PROP: *, POSSESSOR-PHYSOBJ:_),
            CAUSE:_ )
```

where the leftmost symbol L denotes the label of the subframe and the right part is the case expression of the subframe.

The case expression has several pairs of label parts and terms. The label part generally consists of a case label K and a part constructed from several category names C. No specification of categories means that any category can be taken, for simplicity. A single underline denotes a slot of a case, that is, the position of a term to be filled. A double underline shows a slot of the case assigned to the main term of a subframe. In many cases, the main term is the subject term of a text or one of the components. It is common to several subframes and interconnects them to each other.

Table I shows some examples of specification tables for the specific field of semiconductors. Table IA specifies extraction of the general information about semiconductor devices. It consists of the subframes of function property location composition and others.

The FUNCTION subframe specifies extraction of articles that the main term t of the category PRODUCTS takes Physical ACTion to a PHYSical OBJect under a CONDition. The PROPERTY subframe stores the article that a PHYSical QUANTity or a PROPerty of the PRODUCT t has an ATTRibute value or has a RELation to a PARTICipant of a PHYSical QUANTity or a PROPerty of a certain PHYSical OBJect.

Similarly, the subframe label represents the kind of the description about the main term in the subframe. The same

```
(1) (a) "The temperature is increased                    VP6A
          OBJ-PHYSQUANT      PRED-ATTRTRANS
        to 500 °C "
           GO-VALue
    (b) "The drain electrode is connected                VP14
           OBJ-PHYSOBJ          PRED-CONNECTACT
         to the output terminal"
              PARTIC-PHYSOBJ
    (c) "This method is applied                          VP14
         OBJ-THINGS      PRED-USE
           to annealing gallium arsenide"
                PURPOSE-EVENT
(2) (a) "These parameters are used                       VP16A
          OBJ-THINGS        PRED-USE
        to calculate minority carrier lifetime"
                PURPOSE-EVENT
    (b) "This effect  causes                             VP17A
         AG-THINGS      PRED-CAUSE
        the drain conductance to increase"
                OBJ-EVENT
```

Figure 1.

specification is applied to the components of the subject term of a text up to the specified level. Some examples of extracted results are shown later in the example shown in Figure 4.

Table IB is a specification that focuses attention on extraction of the manufacturing process of devices. The specification consists of subframes such as the manufacturing process, the principle, and the composition of the process. The specification is also iteratively used for extraction of subprocesses contained in the subframe COMPOSITION as components. The specification of Table IC pays attention to properties of physical objects under various conditions.

PARSING

Case Structures. Most languages contain a lot of words that have many meanings and many syntactic roles. English is not an exception. Fortunately, however, English has several good dictionaries that describe precise and useful syntactic patterns like Hornby's verb patterns[12] to resolve linguistic ambiguities. Let us consider the example shown in Figure 1.

The "to" prepositional phrase is used syntactically as an obligatory case and also as an optional case for the main verb. In Figure 1–1, it is found from the *Oxford Advanced Learner's Dictionary*[13] that the verb "increase" does not take the verb pattern containing a prepositional phrase while both the verbs "connect" and "apply" take the verb pattern VP14 (vt + direct object + prep. + noun). Hence, the prepositional phrase of example 1a is used as an optional case while those of example 1b,c are used as an obligatory case. Similarly, it is found from the verb patterns of "use" and "cause" that the infinitive of example 2a is adverbial while that of example 2b is complementary.

The authors constructed a case structure for each of about twenty semantic categories of verbs on the basis of Hornby's verb patterns.[15,16] Table II shows some case structures constructed on the verb pattern VP14. The first row shows the syntactic structure consisting of syntactic case labels. The second and succeeding rows show the corresponding semantic case structures for every semantic category of verbs and consist of several pairs of case labels and semantic categories of the

Table II. Case Structures of Verb Pattern VP14

VP14	SUBJ	OBJ	Prep. Phrase
PRED-Physical TRANSfer	AG-THINGS	OBJ-PHYSical OBJect	GO-PHYSOBJ ∪ PHYSLOC
PRED-POSSessive TRANSfer	AG-THINGS	OBJ-PHYSOBJ	RECIP-PHYSOBJ
PRED-Mental TRANSfer	AG-THINGS	OBJ-MENTal OBJect	RECIP-HUM
PRED-CONNECTing ACTion	AG-PHYSOBJ	OBJ-PHYSOBJ	PARTIC-PHYSOBJ
PRED-USE	AG-PHYSOBJ	OBJ-THINGS	PURPOSE-EVENT
PRED-THINKACT	AG-HUM	OBJ-MENTOBJ	COMPARison-MENTOBJ

constituents. In Figure 1, a pair of a case label and a category is assigned to a word or a phrase by referring to case structures of the main verbs in the respective sentences for subsequent processing.

The authors also constructed the case structure of Japanese verbs on the basis of a set of postposition such as Kakujoshis by using almost the same case labels and category system as English to facilitate mechanical translation between both languages.

Adverbial phrases and clauses outside verb patterns constitute optional cases. The case labels are determined by referring to the prepositions or postpositions, the semantic category class of the central word of the phrase and that of the main predicative word modified by the phrase.

The internal form of a noun phrase or clause is represented as follows:

$$t_i(K_0\text{-}C_0{:}t_0, \ldots, K_i\text{-}C_i{:}*, \ldots, K_n\text{-}C_n{:}t_n) \quad (2)$$

where t_i is the modified part consisting of a nominal word and the expression enclosed with parentheses is the modifying part corresponding to the modifying phrase or clause. The symbol "*" denotes the place of t_i in the modifying part.

It can be considered that noun phrases without predicative words arise from ellipsis of some linkage predicates. These noun phrases denote simple binary relations such as "an object and the components" or "an object and the attributes". The internal form is

$$t_i(K_i\text{-}C_i{:}*, K_j\text{-}C_j{:}t_j) \quad (3)$$

and is constructed by referring to the categories of t_i and t_j as well as prepositions or postpositions.

The input sentences of the text are parsed in parallel from the left to the right by a bottom-up method of a kind of modified extended LINGOL.[14] Each new word or phrase is reduced in all possible ways under various syntactic and semantic constraints such as those for case structures of verbs mentioned above, and the possible fragments of the internal expressions are partially constructed. If the new appearing entities reveal that there is no right part combined with a partial internal expression constructed so far, the expression is removed from the set of partial internal expressions.

Utilization of Knowledge in Specific Fields. For resolving complicated dependency relations in a context, it is necessary to utilize the knowledge related to the specific field. The knowledge has structures and can be represented by a form similar to the specification tables. It gives basic and prototype attributes of entities and relations among them already known so far. Associated with specification table of Specification for Extracted Information, these relations are those of disposition of parts in a physical device and temporal or causal relations among procedures in a process.

Table III. A Knowledge Database of Semiconductor Devices: MOSFET

COMPOSITION

OBJ	COMP
MOSFET	{semiconductor substrate, insulating layer, source electrode, gate electrode, drain electrode}
semiconductor substrate	{source, drain, gate, channel}

LOCATION

OBJ	LOC
insulating layer	-on semiconductor substrate
source electrode	-on source
drain electrode	-on drain
gate electrode	-on gate ∪ channel
gate ∪ channel	-between (source, drain)

PROCESS

Order of process	PRED	OBJ	GO	LOC
(1)	thermal oxidation	semiconductor substrate	insulating layer	-on semiconductor substrate
(2)	etching	insulating layer	window	-in insulating layer
(3)	diffusion	impurities	{source, drain}	-in semiconductor substrate
			gate	-between (source, drain)
(4)	evaporation	metal	source electrode	-on source
			drain electrode	-on drain
			gate electrode	-on gate ∪ channel

Table III shows a knowledge database of semiconductor devices of MOSFETs where the PROCESS frame shows the schematic relation: "forming a GOal product on a LOCation by applying the process (PREDicate) to an OBJect." The terms or entities of a subframe are written in generic terms of semantic categories and are related to the entities of the other subframes through a common generic term. In the word dictionary, each term has a pointer to a generic terms of the knowledge database and is related to other terms through the generic term.

In the example shown in Figure 2, it is not clear without fundamental knowledge of semiconductor devices whether the underlined part a depends on either that of b or c. It is found, however, from the word dictionary used here that the linking words "between" in example 1a and "deposited on" in example 2a lead to the subframe associated with OBJect vs. LOCation in the knowledge database. From Table III, the pairs of the underlined parts a and b are identified as the instances of a

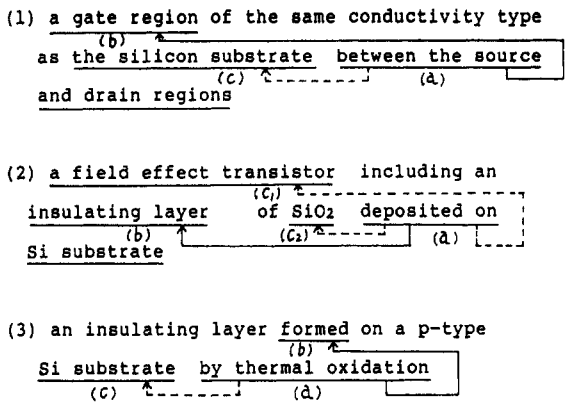

Figure 2.

generic relation between terms of OBJect and LOCation, while a similar relation to these cannot be found between parts a and c. Then it is determined that part a depends on part b.

Similarly, the words "thermal oxidation" in example 3a lead to the subframe of "Process" and are found to depend on an insulating layer formed by the process. If the dependency ambiguities cannot be dissolved by the above method, the system asks the user the correct relation among several candidates.

NORMALIZATION AND INFORMATION EXTRACTION

At the semantic level of internal expressions used here, texts generally have several different internal expressions that describe the same relative relations and properties. These different expressions arise from the free use of synonyms and, furthermore, from the application of different structural expressions depending on the various contexts. In this section, a method of normalization in the latter case is clarified briefly by some illustrative examples. The normal or standard form is set up here to be almost the same form as subframes of specification tables.

These subframes do not involve embedded predicate forms such as "cause to do". They do not contain a kind of copula verbs such as "have" or "consist of" for a binary relation but use a pair of more concrete case labels such as "COMPonent" vs. "OBJect". They also do not contain a term consisting of the same nominal or predicative words as their case labels.

The internal expressions of text obtained by parsing are transformed into the above normal form step by step. First, if the main predicate belongs to a copula-like category, the internal expression is transformed into an apposition form. Then the apposition form is reduced to the abridged form of the subframes by using more concrete case labels.

If an internal expression contains a term consisting of the same words as case labels of the subframes of the specification table, the internal expression is also normalized through an apposition form in a similar way (see note 11). Several examples of normalization are shown in Figure 3.

In the internal expression 1b, the apposition form is reduced to an abridged form of the right side by substitution of the concrete terms in apposition. In the internal expression 2b, the word "include" is found to belong to a kind of copula verb from the word dictionary and replaced by apposition form 2b-2 having a term of "composite", which means "that which includes something". Then the expression 2b-2 is reduced to abridged form 2b-3 and, furthermore, normalized to the relation of OBJect vs. COMPonent of 2b-4.

From the normalized internal expressions, the items specified by the specification table are extracted. First, the main term of a text is identified. In most cases, the main term of a patent

```
The internal expressions of three sentences:
(1a) "Phosphorus, boron and arsenic are materials
     for doping films" ,
(2a) "The semiconductor device includes
     a layer of p-conducting Si",
and
(3a) "The insulating layer is formed by
     oxidizing a Si substrate"
are normalized respectively through the following apposition
forms.
(1b) (OBJ-PHYSOBJ:{phosphorus, boron, arsenic},
     APPOS-PHYSOBJ: materials
       (INSTR-PHYSOBJ:*, PRED-ATTRTRANS:dope,
                  OBJ-PHYSOBJ:films ))
  ⟹(INSTR-PHYSOBJ:{phosphorus, boron, arsenic},
     PRED-ATTRTRANS:dope, OBJ-PHYSOBJ:films )
(2b) (PRED-INCLUSION:include, OBJ-PHYSOBJ:
     semiconductor-device-1,
     PARTIC-PHYSOBJ:layer (OBJ:*, COMP-PHYSOBJ:      (1)
     Si (PRED-ATTR:p-conducting, OBJ-PHYSOBJ:*)))
  ⟹(OBJ-PHYSOBJ:semiconductor-device-1, APPOS-
     PHYSOBJ:composite (COMPOSITE-PHYSOBJ:*,       (2)
       OBJ-PHYSOBJ:layer (OBJ:*, COMP-PHYSOBJ:
         Si (PRED-ATTR:p-conducting,
           OBJ-PHYSOBJ:*))))
  ⟹(COMPOSITE-PHYSOBJ:semiconductor-device-1,
     OBJ-PHYSOBJ:layer (OBJ:*, COMP-PHYSOBJ:       (3)
       Si (PRED-ATTR:p-conducting,
         OBJ-PHYSOBJ:*)))
  ⟹(OBJ-PHYSOBJ: semiconductor-device-1,
     COMP-PHYSOBJ: layer (OBJ: *,                  (4)
       COMP-PHYSOBJ: Si (PRED-ATTR:
         p-conducting,
           OBJ-PHYSOBJ: * )))
(3b) (PRED-PRODuction: form, OBJ-PHYSOBJ:
       insulating-layer-1,
     MEANS-ACT: (PRED-ATTRTRANS: oxidize,
       OBJ-PHYSOBJ: Si-substrate-1))
  ⟹(OBJ-PHYSOBJ: insulating-layer-1,
     APPOS-PHYSOBJ: goal(GO-PHYSOBJ: *,
       PRED-ATTRTRANS: oxidize, OBJ-PHYSOBJ:
         Si-substrate-1))
  ⟹(GO-PHYSOBJ: insulating-layer-1,
     PRED-ATTRTRANS: oxidize,
     OBJ-PHYSOBJ: Si-substrate-1)
The symbol "⟹" denotes transformation of the internal
expression by a coversion rule.
```

Figure 3.

claim sentence is the governor of the first noun phrase or clause while that of a technical paper is the main object of a predicative word of research or proposal appearing in a summary or a title. Then the system picks up the modifying subframes of the main term, identifies the labels of the subframes by referring to the category and the case structure of the main verb, and extracts the specified items from the normalized internal expressions. This process continues until the end of the text and is repeated several times until all the specified information is extracted up to the designated level of components or subprocesses.

The example shown in Figure 4 shows some illustrative examples of extraction from technical papers on semicon-

```
(1) (a) Input abstract
"A MIS semiconductor device is fabricated by depositing
gate-insulator and gate-electrode films on a semiconductor
substrate, forming a mask for patterning the gate-electrode
film, forming gate electrodes, implanting ions into source
and drain regions using the mask, and depositing a
phosphosilicate glass."
    (b) Extracted items
(i) level 1;
  PROCESS; (PRED-PACT: fabricate, OBJ-PRODUCTS:
           MIS-semiconductor-device )
  COMPOSITION; (OBJ-PACT: fabricate, COMP-PACT:
              {deposit-1, form-1, form-2,
               implant, deposit-2 }  )
(ii) level 2;
  PROCESS; (PRED-PACT: deposit-1, OBJ-PRODUCTS:
           {gate-insulator-film,
            gate-electrode-film},
            LOC-PHYSOBJ: semiconductor-substrate  )
           (PRED-PACT: form-1, OBJ-PRODUCTS: mask)
           (PRED-PACT: form-2, OBJ-PRODUCTS:
                              gate-electrodes)
           (PRED-PACT: implant, OBJ-PHYSOBJ: ions,
                GO-PRODUCTS: { source-region,
            drain-region}, INSTR-PHYSOBJ: mask )
           (PRED-PACT: deposit-2, OBJ-PRODUCTS:
                      phosphosilicate-glass )
(2) (a) Input abstract
"The resistance of thin insulator films was analyzed. The
resistance varies linearly with insulator thickness and
depends significantly on the work function of the materials.
Introducing traps in the insulator increases the resistance.
The trap energy level greatly affects the resistance."
    (b) Extracted items
  ANALYSIS; (PRED-THINKACT: analyze,
            OBJ-PHYSQUANT: resistance (OBJ-PHYSQUANT: *
            POSSESSOR-PHYSOBJ: thin-insulator-films))
  PROPERTY; (PRED-ATTR ∪ REL: vary,
            OBJ-PHYSQUANT: resistance,
            PARTIC-PHYSQUANT: thickness
             (OBJ-PHYSQUANT: *,
              POSSESSOR-PHYSOBJ: insulator),
            DEGR: linearly  )
           (PRED-ATTR ∪ REL: depend,
            OBJ-PHYSQUANT: resistance,
            PARTIC-PHYSQUANT: work-function
              (OBJ-PHYSQUANT: *,
               POSSESSOR-PHYSOBJ: materials),
            DEGR: significantly )
           (PRED-ATTR ∪ REL: increase,
            OBJ-PHYSQUANT: resistance,
            CAUSE: introducing-traps )
           (PRED-ATTR ∪ REL: be-affected,
            OBJ-PHYSQUANT: resistance,
            CAUSE: trap-energy-level,
            DEGR: greatly )
where 'be-affected' means the form of a passive voice.
```

Figure 4.

ductors. One of the papers is associated with the manufacturing process, and the extracted information of Table IB is shown, where several subprocesses composing the fabrication are extracted at the second level. The other paper is mainly related to the properties of thin insulator film. The extracted information is based on Table IC, where the extracted terms are not replaced by descriptors or keywords for simplicity.

A similar experiment of information extraction from texts written in Japanese has been also done.[16,17] The extracted information can be written in the specified language of English or Japanese, irrespective of the source language used in the text, by term-by-term translation of the extracted information written in the specification form.

CONSTRUCTION OF A DATABASE

The extracted information of texts in a form of the specification table is transformed to relational tables or inverted files. A relational table is constructed for each subframe of the specification table. Each case label in a subframe is inherited as an attribute name of a relational table.

In a subframe, optional cases and their values such as INSTRument are not always contained in a document. Hence, these cases are separated from the main relational table in order to reduce the useless memory and are recorded in an auxiliary table. An attribute or an item that designates the subject term of each document can be also added to the relational table by users options.

An inverted file is also constructed for the specification table. Each descriptor is chosen from terms appearing in the specification table. It has both the subframe label and the case label of the descriptor beside the document number, and furthermore, by option, it can take an item that shows whether the descriptor is the subject term of the document or not. The transformation of information in a specification table to the above forms of the database can be performed without difficulty.[10]

CONCLUSIONS

The method described above will be useful for extracting and processing various forms of structured information of text in a unified way. The system is written in the LISP language. The required memory for processing is almost 320K bytes, and the average processing time from parsing to storing into relational tables is about 50 s in an interpreter mode of the ACOS 700 for a text sample of about 100 words.

Text generally contains figures, tables, graphs, and other types of illustration besides sentential information. It is desired and can be expected that these types of information will also be reduced to almost the same internal or intermediate expressions as sentential information and handled to augument the sentential information of a text at a semantic level in the near future.[17]

REFERENCES AND NOTES

(1) Salton, G. "Dynamic Information and Library Processing"; Prentice-Hall: Englewood Cliffs, NJ, 1975.
(2) Edmundson, H. P. "New Methods in Automatic Extracting". *J. Assoc. Comput. Mach.* **1969**, *16* (2), 265–285.
(3) Charniak, E. "The Case-Slot Identity Theory". *Cognit. Sci.* **1981**, *5*, 285–292.
(4) Minsky, M. "A Framework for Representing Knowledge"; Winston, P., Ed.; McGraw-Hill: New York, 1975.
(5) Rush, J. E.; Salvador, R.; 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.* **1971**, *22* (4), 260–274.
(6) Rumelhart, D. E. "Notes on a Schema for Stories"; Bobrow, D. G.; Collins, A., Eds.; Academic Press: New York, 1975; p 211.
(7) Hobbs, J. R. "Coherence and Interpretation in English Texts". *International Joint Conference on Artificial Intelligence, 5th* **1977**, 110.
(8) Kinukawa, H. "Japanese Sentence Analysis for Information Retrieval System". *Inf. Process. Soc. Jpn.* **1979**, *20* (10).

(9) Nishida, F.; Kishimoto, G.; Takamatsu, S. "Extraction of Items from Abstracts". *International Joint Conference on Artificial Intelligence, 6th* **1979**, 656.
(10) Takamatsu, S.; Fujita, Y.; Nishida, F. "Normalization of Titles and Their Retrieval". *Inf. Process. Manage.* **1980**, *16*, 155.
(11) Nishida, F.; Takamatsu, S. "Structured Information Extraction from Patent-Claim Sentences". *Inf. Process. Manage.* **1982**, *18* (1), 1.
(12) Hornby, A. S. "Guide to Patterns and Usage in English"; Oxford University Press: Oxford, 1975; 2nd ed.
(13) Hornby, A. S. "Oxford Advanced Learner's Dictionary of Current English"; Oxford University Press: Oxford, 1980.
(14) Tanaka, H.; Sato, T.; Motoyoshi, F. "Predictive Control Parser: Extended LINGOL". *International Joint Conference on Artificial Intelligence, 6th* **1979**, 868.
(15) Nishida, F.; Takamatsu, S.; Kuroki, H. "English-Japanese Translation through Case-Structure Conversion". *International Conference on Computational Linguistics, 8th* **1980**, 447.
(16) Nishida F.; Takamatsu, S. "Japanese-English Translation through Internal Expressions". *International Conference on Computational Linguistics, 9th* **1982**, 271.
(17) Nishida, F.; Takamatsu, S.; Fujita, Y. "Extraction of Structured Information from Texts with Various Forms of Information". *International Conference on Text Processing with a Large Character Set* **1983**, 256.

Prediction of Product Quality from Spectral Data Using the Partial Least-Squares Method

ILDIKO E. FRANK,[†] JOHN FEIKEMA,[‡] NICK CONSTANTINE,[‡] and BRUCE R. KOWALSKI*[†]

Laboratory for Chemometrics, Department of Chemistry BG-10, University of Washington, Seattle, Washington 98195, and 3M Center, Engineering Systems and Technology Laboratory, St. Paul, Minnesota 55144

Received August 8, 1983

Complex regression models for product quality control are calculated by partial least squares with latent variables. Predictor variables from different spectral sources are treated as separate blocks, combinations of which, according to a preset pathway, give prediction models for quality-control variables of an industrial product.

INTRODUCTION

Quality control of raw materials, intermediates, and final products is one of the crucial points in industrial processes. It is the task of analytical chemistry to provide sufficient information on which to base decisions allowing correction of the processes. There are a broad variety of data processing methods applied to quality-control problems in analytical chemistry.[1]

In the 1970s, H. Wold developed a method[2] commonly called partial least squares or PLS referring to a solution for the regression model. This method was developed for application in social sciences but, in the last few years, has found application in chemistry as well.[3-7] PLS is primarily designed for causal-predictive analysis of complex problems that are rich in data but scarce in theoretical knowledge. The basic idea is to separate predictor variables, coming from different influence sectors or sources, into blocks and describe each block by a set of latent variables. These latent variables are linear combinations of the original predictor variables, mutually orthogonal in the same block, and the ones from different blocks correlate to each other according to a preset causal path model.

In the following, a part of a complex study aimed at building predictive models by PLS for product quality control and process control problems is discussed. Our purpose is to point out the potential of the PLS method for use in quality-control problems and its advantages over other multivariate statistical methods. The requirements of proprietary confidentiality do not allow a description of all the data in detail. However, this is not a limitation to understanding the philosophy of the PLS method, illustrating its performance, and discussing the advantages of its applicability in quality control.

METHOD

Multiple linear regression by ordinary least squares is a well-known solution to the problem of describing the variance of a response variable (e.g., a quality-control variable) by a linear combination of predictor variables.[8] The ordinary least-squares algorithms, however, handle only one block of predictor variables and one response variable. In the actual practice of quality control, it is possible to measure predictor variables from different sources or different influence sectors. In this case, different chemical measurements should be handled separately in different blocks or matrices. Another version of the regression problem is the case where there is not only one but several response variables to be described and predicted together by the model. In case of an underdetermined system (more predictor variables than samples), the ordinary least-squares solution incorporates only a limited number of predictor variables in the regression model. The PLS method was developed for the case where several influence sectors are present, each represented by a separate block of variables and connected by a predetermined causal pathway. This method can handle several predictor blocks, multiple response variables, and underdetermined systems. In the following, three special cases of PLS used in this study are described.

(A) PLS Regression. One special case of PLS is distinguished by having one set of predictor variables and only one dependent variable. This is a normal multiple regression case but solved by the partial least-squares algorithm.[9] This algorithm is analogous to principal component regression as it attempts to span the block of the original predictor variables by orthogonal latent variables, which are linear combinations of the original variables. However, there is a difference in the criteria for calculating these latent variables; namely, they describe only the amount of variance in the original variables predictive for the response variable. The response variable is then modeled by a linear combination of these orthogonal latent variables. The maximum number of the latent variables NL_{max}, similar to principal components, is the number of the original variables NV. The optimal NL is found by cross validation,[10] a method that compares the predictive power of the models and chooses the one with the optimal number of latent variables for prediction. Due to the orthogonality of latent variables, the method gives a solution to the collinearity

[†] University of Washington.
[‡] 3M Center.

0095-2338/84/1624-0020$01.50/0