1970 CSI Data—176,000 WLNs

WLN Symbols	Freq.	WLN Symbols	Freq.	WLN Symbols	Freq.
L E5 B6	4990	N HNJ	1157	NUNR D	754
T6OTJ	4376	O- BT6	1148	L6UTJ A	749
-SI-1&	3246	OV1 DO	1094	SWR D&	745
T56 BN	2870	O1 EO1	1078	NR CNW	742
L6TJ A	2772	V1 DOV	1071	T56 BOJ	738
T5OTJ	2767	L50J O-	1048	WNR CN	733
T6NJ B	2542	OV1 EO	1040	L66J C	714
L66J B	2363	T6N CN	1029	L5OJ A	713
UTJ A E	2336	N CN E	1024	T5SJ B	710
Q DQ E	2322	T5NN D	1021	WR D&	698
T66 BN	2040	VTJ A	1019	PO&O2&	685
O1 DO1	1890	UNMR B	1015	T5NNJ A	670
N CNJ	1781	L3TJ A	986	T56 BM	660
T6NTJ A	1772	MR BNW	982	-FE--	655
OTJ CQ	1771	N DNJ	978	N DNTJ	651
T C666 B	1763	T5OJ B	967	O EVJ	648
OTJ B1	1758	OSWR D	964	O GOTJ	632
N DOTJ	1638	T B656	963	T66 BM	631
T56 BN	1631	V1 EOV	959	VNVJ C	630
T6N DO	1578	N FN H	959	T6N DNTJ	630
N DN F	1529	T5O CO	955	T3OTJ	607
T56 BM	1493	OTJ CO	920	NUNR B	601
Q EQ F	1486	T B666	898	OV1 F1	597
N ENJ	1463	O-SI-1&	881	V BUTJ	592
T66 BO	1446	NMR BNW	878	T55 BO D	591
T6N CNF	1416	T66 BN	857	G DG E	590
OTJ BO	1366	PR&R&R	831	C-14 &	588
O COTJ	1302	O1 IO1	831	T6NJ C	587
NW DNW	1229	T6NVMV	824	UTJ B	581
T56 BV	1212	OPQO&O	813	M DNJ	581
OTJ B	1198	T56 BO	796	NTJ A1	580
T66 BO	1187	OTJ C	780	O DO G	576
T66 BV	1163	T5NTJ A	763	T5N CS	567
				NTJ AV	565

Figure 11. One-hundred fragments (6 or more symbols) appearing most frequently in 1970 *CSI*

searches which heretofore were impossible. The advantages of using CSI are both short- and long-range. The short-range advantages are demonstrably economic and include reduced product development costs, shorter research time, and avoidance of duplicative research. The long-range advantages, while less obvious, are hardly less important. The history of research and development shows that chemical discoveries have led to the introduction of new industry or great expansion of existing technologies. The use of CSI by those engaged in research and development tasks should aid in making similar advances.

LITERATURE CITED

(1) Sorter, P. F., Granito, C. E., Gilmer, J. C., Gelberg, A., and Metcalf, E. A., "Rapid Structure Searches via Permuted Chemical Line-Notations," *J. Chem. Doc.* **4**, 56-60 (1964).

(2) PB 180 901, "Wiswesser Line Notations Corresponding to Ring Index Structures," Chemical Abstracts Service, distributed by Clearinghouse for Federal Scientific and Technical Information, Springfield, Va. 22151.

(3) "The Ring Index, Second Edition," Chemical Abstracts Service, Columbus, Ohio 43210.

(4) Granito, C. E., Becker, G. T., Roberts, S., Wiswesser, W. J., and Windlinx, K. J., "Computer-Generated Substructure Codes (Bit Screens)," *J. Chem. Doc.* **11**, 106-110 (1971).

(5) Garfield, E., Revesz, G. R., Granito, C. E., Dorr, H. A., Calderon, M. M., Warner, A. W., "Index Chemicus Registry System: Pragmatic Approach to Substructure Chemica Retrieval," *J. Chem. Doc.* **10**, 54-8 (1970).

(6) Revesz, G. S., Granito, C. E., and Garfield, E., "One-Letter Notation for Calculating Molecular Formulas and Searching Long-Chain Peptides in the Index Chemicus Registry System," *J. Chem. Doc.* **10**, 212-16 (1970).

Syntactical Proximity—Partial Syntactical Analysis of Natural Language Data Records

TAKEO YAMAMOTO* and SHIZUO FUJIWARA

Department of Chemistry, Faculty of Science,
The University of Tokyo, Hongo, Tokyo 113, Japan

Received May 13, 1971

A definition is given for the syntactical proximity of a string in a natural language data record. It uses two dictionaries for the "delimiters" and a limit of length n_{prox} in the search for the proximity. It is shown that, by giving suitable entries for the dictionaries, several kinds of partial syntactical analyses may be performed by a relatively simple operation.

Information retrieval operations using natural language data bases such as CAS data tapes are mostly done by the term-match method. The query provides the system with a dictionary of meaningful terms—words, phrases and/or fragments thereof. A record is retrieved whenever it contains the required combination of the terms. To retrieve records containing information about iron complexes, for example, one provides a dictionary consisting of two sets of terms ("parameters")—one with terms such as "ferrous ", "ferric " and "iron", and the other, with terms

* To whom inquiries should be sent.

such as "complex" and "coordination". If the record contained at least one term from each parameter, it is assumed that the record is relevant. However, this is valid only so long as the record is fairly short and simple, such as titles of journal articles. It would not work with abstract texts, much less with full texts. Even in the case of searching title data, broad terms such as "analysis", "complex", and "determination" tend to be avoided as the search terms because they become too vague if taken out of context, and are thought to be less meaningful than more specific, or narrower, words.

All of this will be different if there is a way of at least partially analyzing the syntactical relationship between the words when they occur in the data records. If, for example, it is known whether a string "iron" is contained in an adjective phrase of a word containing the string " complex", one can be almost sure that the record is relevant to one's interest in iron complexes.

In the present paper, a method will be proposed which enables one to perform the above kind of syntactical analysis of natural language data. Unlike most of the existing methods for syntactical analysis,[1-4] it is simple and may conveniently be incorporated in large-scale information retrieval systems.

DEFINITIONS

A string, M, and a natural number, n_{prox}, are assumed to be given by the query. Two dictionaries, the preceding delimiter dictionary and the following delimiter dictionary (hereafter collectively called the delimiter dictionaries), are also assumed to be given.

A preceding (following) delimiter is a string which is an entry of the preceding (following) delimiter dictionary.

The preceding (following) proximity of M in a record is a string which satisfies all of the following conditions:

1. It occurs in the record.
2. It is equal to or shorter than n_{prox} in length.
3. It is immediately followed (preceded) by M.
4. Unless it is equal to n_{prox} in length or its beginning (end) coincides with the beginning (end) of the record, it is immediately preceded (followed) by a preceding (following) delimiter.
5. No preceding (following) delimiter occurs in it.

The syntactical proximity, S_{m}, of M is the logical sum of the preceding proximity and the following proximity of M.

DISCUSSION

The syntactical proximity of M, S_{m}, as defined above depends on the number n_{prox}, the delimiter dictionaries and the term M. Once the above data are given, an algorithm for finding S_{m} in a record should be straightforward. Some illustrative examples are given in the following:

Example 1. If one gives " " as the only entries in both of the delimiter dictionaries and gives M with a blank at each end, then S_{m} consists of the words next to M. (On the other hand, if one gives M truncated at both ends, then S_{m} consists of the rest of the word.)

Example 2. If no entry is given for the delimiter dictionaries, then S_{m} will depend only on n_{prox}, that is, it coincides with the physical proximity of M.

Example 3. If one gives lists of sentence terminators, verbs, prepositions, relative pronouns and auxiliary verbs as the preceding delimiter dictionary, and lists of sentence terminators, verbs, relative pronouns and auxiliary verbs as the following delimiter dictionary, S_{m} (of a noun M) corresponds approximately to the adjective phrase of M. (Sentence terminators, the delimiters used for terminating sentences, depend on the data to be analyzed. In the printed form of *Chemical Abstracts,* for example, sentences are terminated by ". ", whereas in other technical writings delimiters such as ".) ", ". " and "? " may be used as sentence terminators.)

As shown above, the present method may be applied to a wide variety of syntactical analysis. As the method is relatively simple and fast, it may be used in processing large natural language data-containing abstracts and texts. The simplicity of the method comes from the fact that, given strings M and N, we already know much about the possible syntactical relationship between them. Combined with the knowledge, even a limited syntactical analysis such as the present one may be very effective. In this connection, it should be pointed out that the delimiter dictionaries do not have to be exhaustive for the method to work. For example, when many verbs in the data records are preceded by auxiliary verbs (which is the case in most scientific writings), the list of verbs in Example 4 may be omitted from the dictionaries, greatly simplifying the analysis.

An information retrieval system is now being built at the University of Tokyo, using the partial syntactical analysis described above.

LITERATURE CITED

(1) Sager, N. N., "Syntactic Analysis of Natural Language," in *Advances in Computers,* F. L. Alt and M. Rubinoff, Eds., Vol. 8, p. 153, Academic Press, New York, N. Y., 1967.
(2) Salton, G., "Automatic Information Organization and Retrieval," Chap. 5, McGraw-Hill, New York, N. Y., 1968.
(3) Bobrow, D. G., "Natural Language Input for a Computer Problem-Solving System," in "Semantic Information Processing," M. Minsky, Ed., MIT Press, Cambridge, Mass., 1968.
(4) Abe, N., Toyoda, J., and Tanaka, K., "Some Considerations on an Automatic Indexing System by Use of a Title of the Document," *Joho Shori* 11, 699 (1970).