An Approach to Computational Semantics of Natural Languages

HIROHIKO NISIMURA* AND SHUICHI IWATSUBO*

In this paper a new approach to computational semantics of natural language is presented.

We assume that a sentence consisting of a set of words is an entity capable of transmitting information. Some words, later called sample words are extracted from the text. In Fig. 1 the cooccurence distribution of sample words and

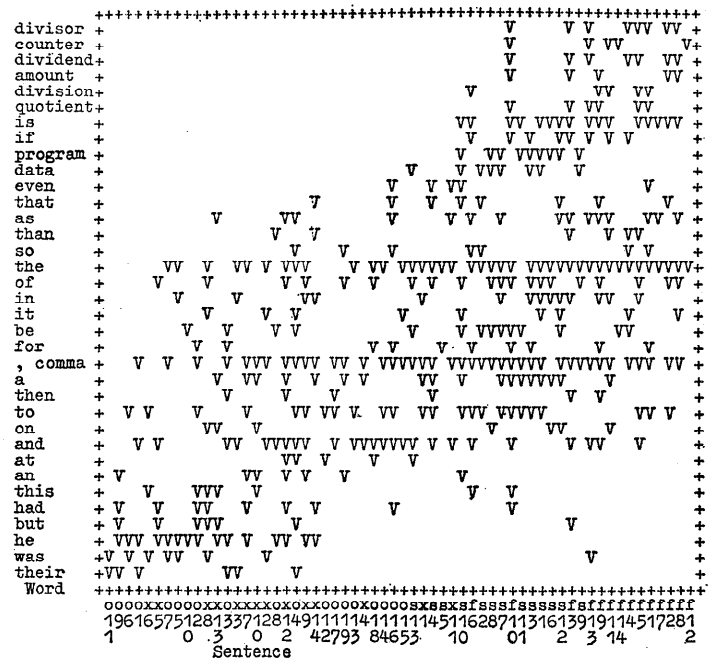

Fig. 1. Joint Distribution Matrix of Words and Sentences.

sentences is shown. Using this table we can group the sample words according to the degree of similarity in the sentences. We can mathematically realize this procedure in the following way. With each sample word and with each sentence is associated a real number. n sample words are represented by an n-dimensional vector, each component corresponding to the numerical value of a sample word. Those numerical values are determined in such a way that the

This paper first appeared in Japanese in Joho-Shori (the Journal of the Information Processing Society of Japan) Vol. 11, No. 3 (1970), pp. 127-134.

* Electrotechnical Laboratory, MITI, Tokyo.

correlation coefficient between sample words and sentences is maximized. This leads to the equation $A\boldsymbol{x}=\lambda B\boldsymbol{x}$, where A is a real symmetric matrix whose elements A_{ij} (i, $j=1\sim n$) represent the degree of similarity between the i-th and j-th sample words and B is a diagonal matrix whose elements are the frequency of the sample words. Solving this eigenequation we obtain a set of eigenvalues λ and the corresponding eigenvectors $\boldsymbol{x}$. Each eigenvalue is equal to the square

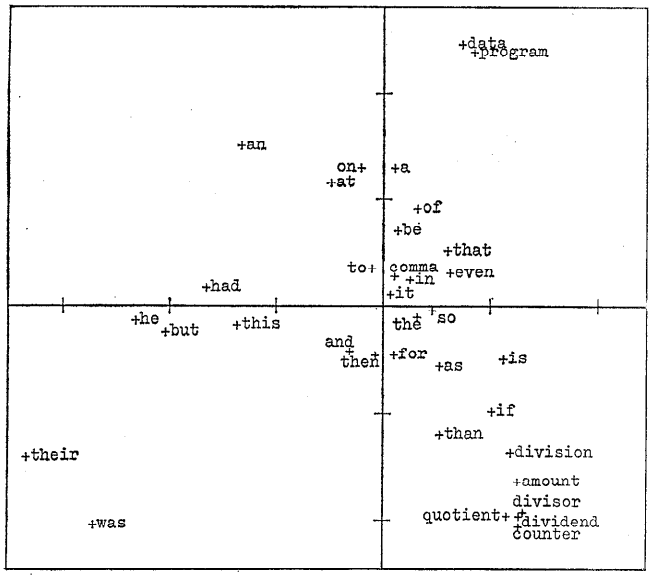

Fig. 2. Configuration of Words in Semantic Space.

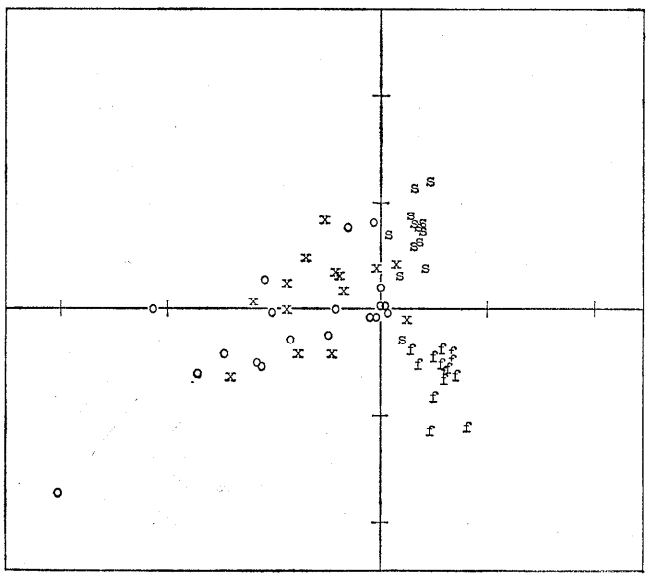

Fig. 3. Configuration of Sentences in Semantic Space.

of the correlation coefficient and the components of the corresponding eigenvector represent the numerical values associated with each sample word. The numerical value corresponding to a sentence is given by the average value of the sample words contained in this sentence.

The result of an experiment using a small number of sample words and sentences is illustrated in Figs. 2 and 3. The text used in the experiment was taken from scientific papers and from some modern novels shown in table 1. The frequency of the extracted sample words is shown in Table 2.

Table 1. Sampled Sentences.

Group	Subject	Number of Sentence	Running Words	Different Words
s	Computer software	13	291	247
f	Computer hardware	14	375	293
o	Steinbeck's novel	19	285	261
x	Bellow's novel	14	303	259
Total		60	1,254	483

Table 2. Frequency of Sample Words.

	Word	Freq.		Word	Freq.		Word	Freq.
1	comma	41	13	had	9	25	so	7
2	the	40	14	it	9	26	this	7
3	and	25	15	program	9	27	quotient	6
4	to	24	16	data	8	28	than	6
5	of	20	17	divisor	8	29	their	6
6	a	16	18	if	8	30	then	6
7	is	16	19	that	8	31	amount	5
8	as	15	20	was	8	32	at	5
9	he	15	21	an	7	33	counter	5
10	be	14	22	but	7	34	division	5
11	in	14	23	dividend	7	35	even	5
12	for	10	24	on	7			

In Fig. 2 the sample words corresponding to the two largest eigenvalues are represented, while the sentences corresponding to those eigenvalues are represented in Fig. 3. The distance between two points in Fig. 2 or 3 is proportional to the degree of similarity between sample words or respective sentences.

High frequency words are located near the origin of coordinates. Words and sentences from the computer field are located in the righthand side of diagram 2 and 3 while those from modern novels are located in the lefthand side. Furthermore, words and sentences from the software-section of the scientific papers are located in the upper righthand side and those of hardware-section in the lower righthand side.

This method may also be applicable to automatic classification of documents and construction of a thesaurus.