Extraction of Semantic Relationships among Terms to Construct Organized Knowledge Resources

Takayuki Morimoto[†], Tetsuya Maeshiro[‡], Yuzuru Fujiwara[†]
[†] Department of Information Science, Kanagawa University
[‡] ATR Human Information Processing Research Laboratories

Abstract

While the global borderless flow of diversified huge amount of information is being developed at unprecedented speed, major functions of information processing i.e. numerical calculation, symbol matching in information retrieval and deduction, remain to be unchanged. However, requirement and necessity of advanced utilization of contents of information have been recognized gradually. Learning and thinking are worth a while targets to such requirement and have been widely studied without useful results thus far. It is necessary to know meanings and characteristics of terms and various relationships among them in order to realize machine learning and thinking, because technical terms are the most convenient and powerful representation medium of abstract concepts. Therefore, the methods of constructing organized knowledge resources are based on extracting semantic relationships among terms.

1 Introduction

Information technologies are being developed at unprecedented speed due to high performance and inexpensive computers and Internet have been widely available. The transmission and utilization of information become more diversified and borderless very rapidly. However, major functions of information processing by conventional computers have been unchanged, i.e. numerical calculation, symbol matching in information retrieval and deduction. Therefore, the advanced utilization of contents of information are necessary in addition to above-mentioned functions. Machine learning and thinking are the typical targets of such utilization.

Semantic processing and understanding are required to realize learning and thinking. It is necessary to know and to integrate concepts and semantic relationships in a huge amount of stored information. The above requirements are solved by semantic analysis of information and the structuralization of organized knowledge resources based on their attributes, characteristics, meaning, and so on. A model by which multiple hierarchical, overlapping, n-ary, dynamic and relative relationships can be described is devised in order to represent such semantic structures. It is apparent that neither graph model nor hyper graph model has sufficient capability to represent such conceptual structures. The Homogenized Bipartite Model has been proposed so as to satisfy such requirements.[1]

2 Constructions of Organized Knowledge Resources

Generally, thesaurus, taxonomy, or access file has been used in order to make information adapted for managements into knowledge resources. Relationships are divided into three types in these methods, that is, physical, conceptual and logical one called physical, conceptual, causal structures respectively. Physical structures represent physical origin and storage address. Conceptual structures represent conceptual relationships, i.e. hierarchical and other associative relationships. And, causal structures represent various logical relationships including cause/effect relationships.

Especially, thesaurus is a conventional method to represent conceptual structures, and there are many studies as follows:

- thesauri which are constructed manually[2]
- thesauri which are constructed automatically
  - compiling individual relationship for thesauri using collected documents[3]
  - merging two or more thesauri[4]
  - expert system base (dynamic methods using user information)[5]

However, out goal is to realize machine learning and thinking, and these thesauri are not sufficient as far as to consider contents of information.

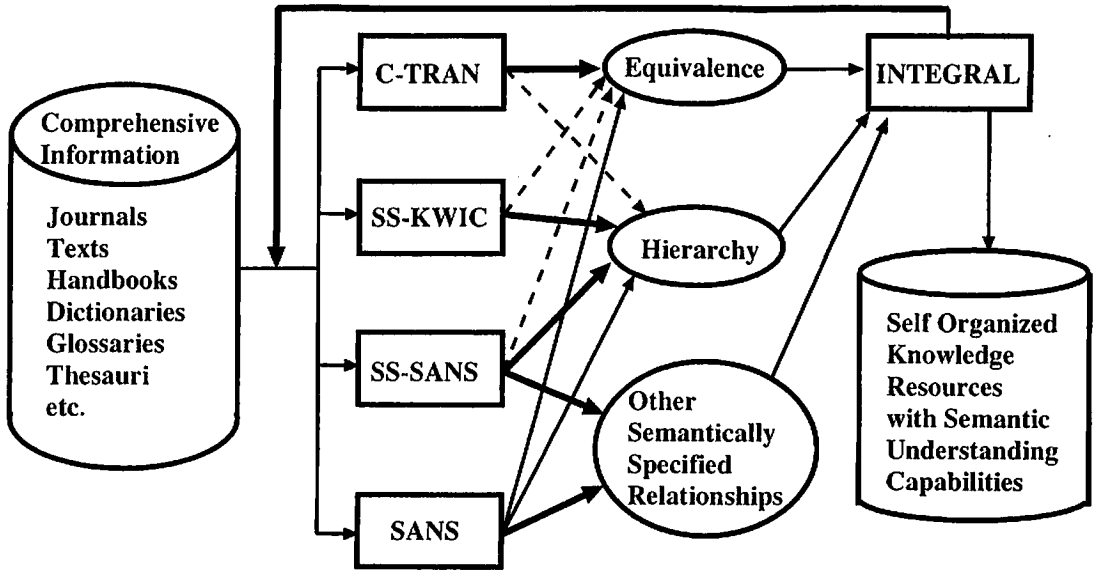

Figure 1: Self organized knowledge resources based on semantic relationships

Figure 1 shows our proposed method for the construction of organized knowledge resources. In our method, equivalent, hierarchical, and various semantic relationships are extracted and integrated to construct thesauri automatically.

By using C-TRAN (Constrained Transitive Closure) based on bilingual relations in glossaries, equivalent (synonym) and hierarchical relationships (terms represented a super-ordinate and a subordinate concepts) can be extracted. [6][8] Hierarchical and associative relationships can be extracted by using SS-KWIC (Semantically Structured Key Word elements Index in Terminological Context) based on modified relations in terminological contexts.[7][8] SS-SANS (Semantically Specified Syntactic Analysis of Sentences) based on syntactic templates and definitions of terminologies can extracts various semantic relationships.[9][10] The function of SANS (Semantic Analysis of Sentences) is semantic analysis of contents. The general structure of a conceptual can be built by using INTEGRAL.

3 Homogenized Bipartite Model

The Homogenized Bipartite Model (HBM) was developed in order to describe semantic relationships between conceptual structures. HBM is an extended Hypergraph, and recursive and nested structures can be described in this model. Relations which can be represented by HBM and conventional graph models in Table 1.

HBM is formulated as follows:

$$E \subseteq 2^V \quad (1)$$

$$V = V \cup E \quad (2)$$

$$E = E \cup V \quad (3)$$

$$L \rightarrow E \cup V \quad (4)$$

V : *a set of vertices*

E : *edges*

L : *labels*

Table 1: Comparison between graph models

Models	Relations				
	Binary	Many-to-many	Overlap	Nest	Relative
Graph	o	×	×	×	×
Hyper Graph	o	o	o	×	×
Homogenized Bipartite	o	o	o	o	o

Table 2: Potency of information

Type	Model	Potency	Bipartite	
Set	S = V	N	S = (V, $\emptyset$, $\emptyset$)	
Tree	S = (V, E)	2N	S = (V, E, L)	$E \subseteq V \times V$
Graph	S = (V, E)	N^2	S = (V, E, L)	$E \subseteq 2^V$
Hyper Graph	S = (V, E')	2^N	S = (V, E', L)	$E \subseteq 2^{2\cdots V}$
Homogenized Bipartite	S = (E", E", L)	$2^{2\cdots N}$	S = (V, E", L)	$L \subseteq (V) \times (E, E', E'')$

Table 3: Correspondence between information structures and described semantic relationships

Type of structure	relation	characteristics(semantic relationship)
Set	–	
Tree	binary	classification : hierarchy
Graph	binary	multiple inheritance, etc.
Hyper Graph	n-ary	partial sharing, duality, etc.
Homogenized Bipartite	n-ary	nested structure, modality, relativity, etc.

The formula (1) represents that many-to-many relations can be described, and it is the same with Hypergraph. Recursive and nested structures are allowed by the formulas (2) and (3) respectively. By the integration between the formulas (2) and (3), nodes(V) and links(E) are homogenized. Table 2 shows the potency of representation in HBM and conventional models, and Table 3 shows structures and characteristics of relationships which can be represented in them. (As a matter of course, lower structures in Table 3 include upper structures.) It is clearly that HBM is more capable than other models.

Moreover, conceptual structures based on HBM can be used for major thinking functions such as induction, analogical reasoning, (analogical) abduction. The mechanisms of such functions are as follows:

Let $C = (V, E)$ be the universe of concepts, and $C_r = (V_r, E_r), C_s = (V_s, E_s), C_c = (V_c, E = c)$, where r, s, and c designate reference, sample, and common substructures respectively. The mechanism of induction:

$$\begin{aligned} C_c &\subseteq C_{si} \cap C_{sj} \cap \cdots \cap C_{sn} \cap C, \\ C_{s'} &= (V_{s'}, E_{s'}) = C_r(V_r, E_r), \qquad (5) \\ i.e.\quad V_r &= V_c + \delta V_r, \\ E_r &= E_c + \delta E_r \ \ and \\ V_{s'} &= V_c + \delta V_r, \\ E_{s'} &= E_c + \delta E_r. \end{aligned}$$

The mechanisms of analogical reasoning and (analogical) abduction:

$$\begin{aligned} C_c &\subseteq C_s \cap C_r \\ C_{s'}(V_{s'}, E_{s'}) &= C_r(V_r, E_r), \qquad (6) \\ i.e.\quad V_r &= V_c + \delta V_r, \\ E_r &= E_c + \delta E_r \ \ and \\ V_{s'} &= V_c + \delta V_r, \\ E_{s'} &= E_c + \delta E_r. \end{aligned}$$

4 SS-SANS

4.1 Concept

SS-SANS is the system to extract various semantic relationships from contents. There are one or more sentence structures based on verbs which represents relationships in a content. In many cases, they represent relationships between more than one set of nouns. Therefore, if a relationship between nouns is understood, a different structure whose relationship is equivalent may be detected. The extracting method of SS-SANS is to repeat above operations. By using this method, wide variety of relationships are extracted. In SS-SANS, there are two features by means of semantic analysis of contexts. One is that dynamic conceptual relationships can be extracted. And the other is that it is easy to extract various relationships from new information. By these features, SS-SANS can extract not only equivalent and hierarchical relations, but also important relations between cause and effect which.

4.2 Algorithms

The algorithm of SS-SANS is shown as Figure 2. The each numbers of operations in Figure 2 correspond to follow statements.

1. syntactic analysis
2. addition of information including parts of speech and partitions of words
3. initialization of the template file (setting of the initial template construction)
4. comparison between sentence structures in the template file and the file made by above operation 2. (The keys of comparison is a construction of a sentence)
5. extracting associative relationships which fill the requirement of keys
6. appending new extracted associative relationships to the set of relationships which are detected previously
7. comparison between the template including associative relationships and the file of contexts with various information
8. extracting constructions matching with the template of constructions
9. appending new constructions to the set of constructions which are extracted previously

Operations from No. 1 to No. 3 are initial processes and executed at least once. On the other hand, operations between No. 4 and No. 9 are major operations and repeated until a new relationship is not extracted. In this algorithm, various semantic relationships can be extracted in compliance with changing the initial input construction.

4.3 Experiment

4.3.1 Environment

The data of "NACSIS test collection" is used as the initial input information data in this experiment. In this data, there are many titles and abstracts of paper concerned with AI, which are already executed syntactic analysis by using "JUMAN" which is an user-extensible morphological analyzer for Japanese.[11]

There are 242,869 words without indexes in this data. The initial input construction is "[NN] を 行う [NN]". [NN] represents normal noun. The character "を" is a postpositional particle([S]), and the word "行う" is a verb([V]) and means "to do" or "to execute". By using this construction, associative relationships can be extracted.

4.3.2 Results and Discussions

Table 4 shows numbers of associative relationships and extracted sentence structures. The number of repeats corresponds to frequency of executed operations from No. 4 to No. 9 (Figure 2).

There is no increase between 7th and 8th repeats, and the number of constructions is 73. Namely, there is no new construction which can be extracted by using associative relationships. Moreover, the maximum number of associative relationships is 2251 in this experiment.

The result in table 5 shows that the relation between the numbers of extracted constructions and associative relationships. Over 10 percent of associative relationships can be extracted two or more constructions.

Table 6 shows examples of associative relationships which can be extracted multiple constructions. In this experiment, many constructions and relationships are extracted by using simple associative predicate such as "[NN] [S] [V] [NN]"

Table 4: The numbers of associative relationships and constructions of sentences

The number of repeats	The number of associative relationships	The number of constructions
0	0	1
1	24	2
2	51	7
3	162	21
4	889	41
5	2026	58
6	2174	67
7	2233	73
8	2251	73

Table 5: The number of associative relationships toward the number of extracted constructions

The number of constructions	The number of associative relationships
1	2034
2	179
3	23
4	8
5	3
6	2
7	1
10	1

Table 6: Associative relationships and the number of extracted constructions

Associative relationship	The number of constructions
"構造"[S][V]"知識"	4
"ニューラルネット"[S][V]"モデル"	4
"コスト"[S][V]"仮説推論"	4
"代入例"[S][V]"定理"	4
"問題"[S][V]"解"	4
"共有ストア"[S][V]"待ち行列ネットワーク"	4
"マルチエージェント系"[S][V]"知識"	4
"状況"[S][V]"学習者"	4
"エージェント"[S][V]"知識"	5
"システム"[S][V]"知識"	5
"学習者モデル"[S][V]"概念形成機能"	5
"モデル"[S][V]"アイロニー"	6
"知識"[S][V]"必要"	6
"知識"[S][V]"方法"	7
"対象"[S][V]"問題"	10

(i.e. "[NN] を 行なう [NN]"). However, by using more complex and concrete construction such as "[NN] [S] [NN] [S] [V] [NN]", it is not easy to extract them. For example, only 6 associative relationships shown as follow are detected by the construction as "[NN] と [NN] を 用いた [NN]".

- "物理特徴概念" [S] "構文解析" [S] [V] "技法機械"
- "統合システム" [S] "図面形状認識" [S] [V] "工程設計支援方法"
- "ドット記法" [S] "IS-A 関係" [S] [V] "推論システム"
- "統計知識" [S] "文脈情報" [S] [V] "一般化 LR 構文解析法"
- "ニューラルネットワーク" [S] "ファジィ推論" [S] [V] "各種"
- "機能学習アルゴリズム" [S] "クラスタリング手法" [S] [V] "概念シソーラス"

Therefore, these results shows an importance of the initial input construction, and may be caused by organization or systematization of knowledge has not made progress in information science.

It is a dangerous to think that all extracted relationships have semantic relationships. A relationship have the confidence until it is compared with sentences and other constructions can be extracted. Consequently, associative relationships which can be extracted tow or more constructions are considerably reliable.

5 Conclusion

Requirements and necessity of advanced utilization of contents of information are recognized because the global flow of information is being developed at unprecedented speed. Machine learning and thinking are the typical targets of such requirements, and studies for the realization of these functions is shown. The structuralization of knowledge resources and functions of thinking such as induction, analogical reasoning, and abduction may be implemented by utilizing these structures. Moreover, an example of the functions for the realization of machine learning and thinking is reported. This function called SS-SANS is extracting various semantic relationships from contents. Prototype of functions for the structuralization have been already developed.

References

[1] Y. Fujiwara and Y. Liu, *The Homogenized Bipartite Model for Self Organization of Knowledge and Information*, IFID, 2(1), pp13-17, 1998.

[2] A. Ghose and A.S. Dhawle, *Problems of Thesaurus Construction*, Journal of the American Society for Information Science, 7, pp211-217, 1997.

[3] D. Soergel, *Automatic and Semi-Automatic Methods as An Aid in the Construction if Indexing Language and Thesauri*, Inter. Classif. 1(1), pp34-39, 1974.

[4] R. Forsyth and R. Rada, *Machine Learning-Applications in Expert Systems and Information Retrieval*, England: Ellis Hornwood Series in Artificial Intelligence, West Sussex, 1986.

[5] U. Guntzer, and et al. *Automatic Thesaurus Construction by Machine Learning from Retrieval Sesseions*, Information Processing and Managament, 25(3), pp265-273, 1989.

[6] Y. Fujiwara, *The Model for Self Structured Semantic Relationships of Information and Its Advanced Utilization*, International Forum on Information and Documentation, vol.1 9(2), pp-8-10, 1994.

[7] J. Lai, H. Chen, and Y. Fujiwara, *An Information-Base System Based on the Self-Organization of Concepts Represented by Terms*, Terminology , vol.3 (2), pp313-314, 1996

[8] Y. Fujiwara and J. Lai, *An Information-Base System Based on the Self-Organization of Concepts Represented by Term*, Terminology, Vol.3(2), pp313-314, 1997.

[9] H. Sano and Y. Fujiwara, *Syntactic and semantic structure analysis of article titles in analytical chemistry*, J. Inf. Sci. Principles and Practice 19, pp119-124, 1993.

[10] H. Sano and Y. Fujiwara, *Automatic Assignment of Word Categories for Improved Facet Analysis of Titles and Indexes*, J. Inf. Sci. Principles and Practice 20, pp23-31, 1994.

[11] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html

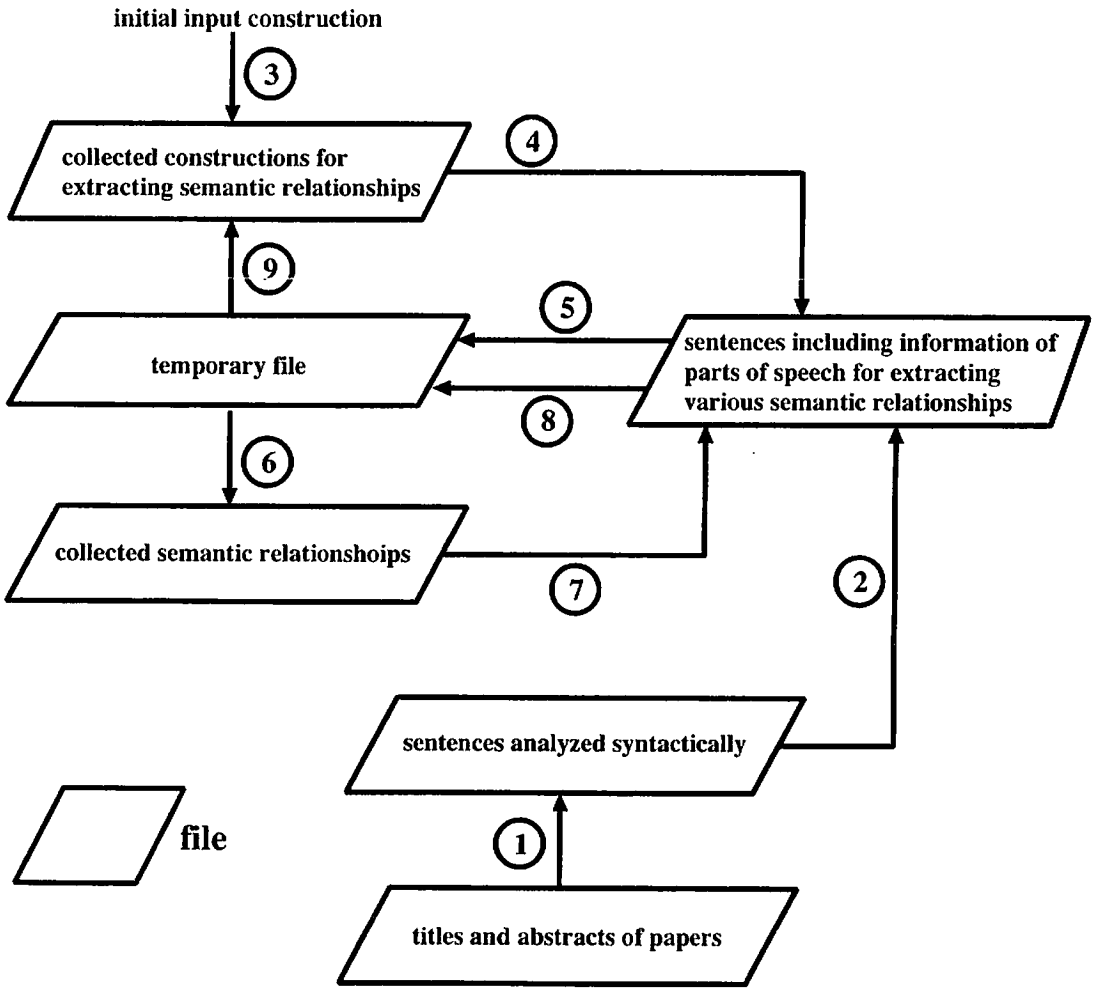

Figure 2: Algorithm of SS-SANS