Automatic Concept-base creation method using document groups

Misako Imono[1], Eriko Yoshimura[2], Seiji Tsuchiya[2], and Hirokazu Watabe[2]
[1] Organization for Advanced Research and Education, Faculty of Science and Engineering, Doshisha University, Kyo-Tanabe, Kyoto, Japan
[2] Dept. of Intelligent Information Engineering and Sciences, Faculty of Science and Engineering, Doshisha University, Kyo-Tanabe, Kyoto, Japan

Abstract - *This paper describe a method for creating concept-bases (CBs), which are knowledge bases comprising concepts that have been mechanically extracted from multiple sources and attributes that express their semantic features. In a CB, concepts are assigned attributes and weightings that express their importance. This means that that it is not necessary to define systematized relationships between concept and attributes, as is the case of a thesaurus, a semantic network, and/or an ontology. Concepts and attributes are defined based solely on the relationships that can be associated with each other. Using such definitions, a CB aims at including various meanings that human beings understand automatically based on words used, not simply definitions as described in dictionaries. The proposed method is capable of automatically building a CB from document groups such as newspaper articles, scientific papers, and Web articles that have not been analyzed in depth. Since this method is not restricted by document type, CBs can be built easily and automatically to suit the intended usage purpose.*

Keywords: Knowledge base, Concept-base, Degree of association, Association mechanism

1 Introduction

Currently, numerous challenges restrict the use of natural languages in information processing technologies in areas such as effective Web searches, recommendation systems, document classification, and robot communication. This is primarily because the natural language information contained in many documents can vary with the addition or subtraction of a single sentence or word. Accordingly, a different approach to defining word or phrase meanings is necessary when dealing with the information contained in natural language.

Human beings understand word, phrase, and sentence meanings flexibly expressed in natural language because their in-depth knowledge allows them to make "meaning" associations outside the definitions appearing in dictionaries. This includes matching contexts based on the other words, phrases, or sentences used. For example, humans readily and naturally find an association between the words "art" and "impression", even though "impression" is not normally included in dictionary definitions of "art".

General natural language processing defines clear relationships as the basic approach to understand word meanings. Thesauruses define the meaning of words by constructions that express super-sub relations and synonyms, while ontologies create models around a certain reality by defining parameters that indicate characteristic and clear relationships with a topic.

However, while significant amounts of knowledge have been systematized and utilized via such techniques, they remain insufficient to make rapid associations based on the example described above. It other words, it would be difficult to express the relationship between the word "art" and the word "impression" utilizing the knowledge that has been systematized in a thesaurus or an ontology. In a thesaurus, the higher node for "art" is "creation", and the higher node for "impression" is "feeling". Common nodes for these words exist only in the "abstract".

While the vague relationships that permit humans to understand complex concept associations cannot be expressed systematically, such association can be identified if the appropriate sources and attributes are included in a concept base (CB)[1], which is a knowledge base consisting of natural word and phrase combinations built by focusing attention solely on their associate relationships.

A CB defines the meanings of various phrases called concepts that are expressed in natural language based on their relationships to other phrases called attributes. This means that neither labels nor categories indicating clear relationships need exist between concepts and attributes. If a source indicates that human beings detect relationships between "art" and "impression", the CB will include "impression" in the attribute of the "art" concept without attempting to define the relationship between them. This CB structure allows meanings to be defined flexibly, much like humans do.

Therefore, in this paper, we describe a method that allows CBs to be built easily and automatically from document groups such

as newspaper articles, academic papers, and Web articles that have not been analyzed in depth. This method is not restricted by document type.

2 Concept-Base

A CB is a knowledge base that defines words as concepts. A concept is defined in the following equation:

$$A = \{(a_1, w_1), (a_2, w_2), \ldots, (a_i, w_i)\}$$

where A is the concept label, a_i is the attribute, and w_i is the weight of the attribute. Table 1 shows specific examples of concepts.

Table 1: Specific example of concepts

Concepts	(Attribute, Weight)
Art	(Masterpiece, 0.34) (Impression, 0.23) (Ceramic Art, 0.12) (Sense of beauty, 0.08) …
Impression	(Sensitivity, 0.18) (Heart, 0.18) (Sense of beauty, 0.04) (Deep, 0.02) …
…	…

An attribute of a concept is called a first order attribute. In the CB, words defined by concepts also form attributes, which can then be used to derive other attributes. Attributes derived from attributes are called second order attributes of the original concept.

Concepts are defined by the synonymous and unforced "associative" relationships. Synonymous unforced relationships exist between the concepts and attributes in CBs. Synonymous relationship are not necessarily clear. However, relationships can be defined based on their "associative" level. This flexible semantic definition is a difference between WordNet.

3 Degree of Association

The Degree of Association (*DoA*)[2] quantifies the relationship between concepts by using attributes that characterize the chain-reaction structure of the CB. Table 2 shows specific *DoA* examples.

Table 2: Specific *DoA* example

Concept A	Concept B	*DoA*
Art	Artwork	0.15
	Impression	0.018
	Routine	0.0015

In this process, the relationship between multiple concepts is expressed quantitatively. The following shows the method used to calculate the *DoA* between Concept A and Concept B. This is defined as *DoA(A, B)*. For concepts A and B with primary attributes a_i and bi, weights u_i and v_j, and numbers of attributes L and M, are respectively ($L<M$), the concepts can thus be expressed as follows:

$$A = \{(a_1, u_1), (a_2, u_2), \ldots, (a_L, u_L)\}$$

$$B = \{(b_1, v_1), (b_2, v_2), \ldots, (b_M, v_M)\}$$

The degree of match (*DoM*) between concepts A and B *DoM (A,B)*, where the sum of the weights of the various concepts is normalized to 1, is defined as follows:

$$DoM(A, B) = \sum_{a_i = b_j} \min(u_L, v_j)$$

The *DoA* is found by calculating the *DoM* for all of the targeted primary attribute combinations, and then determining the relationships between them. Specifically, priority is given to the correspondence between matching primary attributes. For primary attributes that do not match, the correspondence is determined by maximizing the total *DoM*. This makes it possible to give consideration to the *DoA*, even for primary attributes that do not match perfectly. When the correspondences are thus determined, the *DoA(A,B)* between concepts *A* and *B* is as follows:

$$DoA(A, B) = \sum_{i=1}^{L} DoM(a_i, b_{xi}) \times \frac{(u_i + v_{xi})}{2} \times \frac{\min(u_i, v_{xi})}{\max(u_i, v_{xi})}$$

In other words, the *DoA* is proportional to the degree of identity of the corresponding primary attributes, the average of the weights of those attributes, and the weight ratios.

4 Automatic CB creation method

If an information source is defined by a direction-word and sentence pairs that express the meaning of the direction-word, it is possible to easily build a CB by defining the direction-word as the concept and the other words in the sentence as attributes. However, human beings constantly make various kinds of word association that are not found in dictionary definitions, and significant amounts of such information exist in miscellaneous sentences of various documents. In addition, the word knowledge that human beings use when making associations also exists in the miscellaneous sentences they utter naturally.

This paper proposes a method of extracting concepts and attributes automatically from document group information sources that have not been analyzed. Document group examples include newspaper articles, academic papers, and Web articles. Indeed, any document type can be used as an information source as long as it contains sentences that are suitable to the CB usage purpose.

4.1 Information source

In this study, a CB was automatically created using a year's worth of Japanese newspaper[3] issues as an information source. The field was not limited by article type and all newspaper articles, a total of 111,497, were examined.

4.2 Acquisition of concepts and attributes from co-occurrence range

A sentence (which is the range divided by periods within an article) is used to define the co-occurrence range. Concept and attributes pairs are acquired from this range. Figure 1 shows specific examples of co-occurrence range in article.

…Managing Director of xxxx Corporation emphasized the importance of a precise judgment while slump of the economy was whispered. "I think that it is important to investigate a state of the economy through the conversation with the on-site person in charge."

From slump of the demand in the iron manufacture industry, he look in retreat of the economy….

Figure 1: Specific example of co-occurrence range in article.

In figure 1, three sentences extracted from an article are shown. The underlined sentence is an example of the co-occurrence range from which words and phrases such as "economy", "slump", "whisper", "importance", "judgment", among others, are extracted. These words and phrases are defined as concepts and attributes for each. In the above example, the attributes "slump", "whisper", "importance", and "judgment" define the concept of "economy". This process performs such definition of concepts and attributes for all information sources. Figure 2 shows specific concept and attribute definition examples.

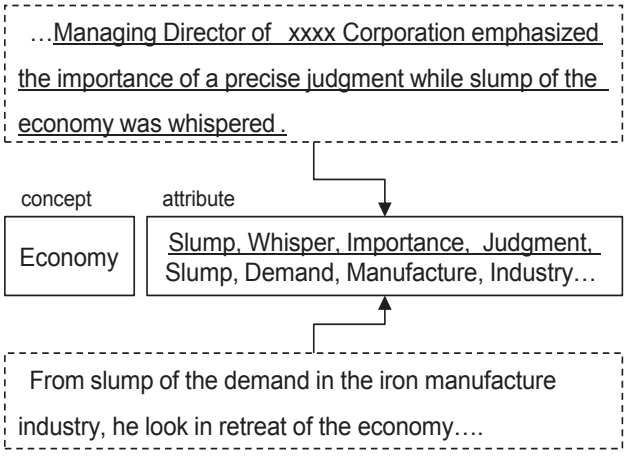

Figure 2: Specific example of definition of concepts and attributes.

The "economy" concept and its attributes are acquired from the first co-occurrence range (underlined portion). Afterwards, attributes such as "slump", "demand", "manufacture", and "industry" are added to the "economy" concept because they appear within the same co-occurrence range. After examining all articles for concept and attribute acquisition, a total of 316,319 concepts were extracted.

4.3 Weighting of attributes

TF-IDF[4], which is a commonly used technique for weighting words in documents for document searches, etc., is used to assign a weight that expresses the importance of each attribute. In this process, *TF* is the term frequency (the appearance frequency of the words) and *IDF* is the appearance inverse document frequency. These products calculate the weight of words. When N pieces of documents exist, the weight of word t which appears in document d is expressed as $TF(t,d)$, which is the frequency of t in document d, and $IDF(t)$ is expressed as shown below:

$$IDF(t) = \log_2\left(\frac{N}{df(t)}\right) + 1$$

where $df(t)$ is the total number of documents in which word t appears.

In attribute weighting, one concept is regarded as one document, and attributes of this concept are regarded as the words in the document. Therefore, the number of concepts (316,319) is regarded as the number of all documents N. The weight of attribute a_i of Concept A is calculated from $TF(a_i,A)$, which is the number of times that attribute a_i is collocated with Concept A, and $IDF(a_i)$ is calculated with the number of concepts that have a_i in its attribute. Figure 3 shows specific weighting examples.

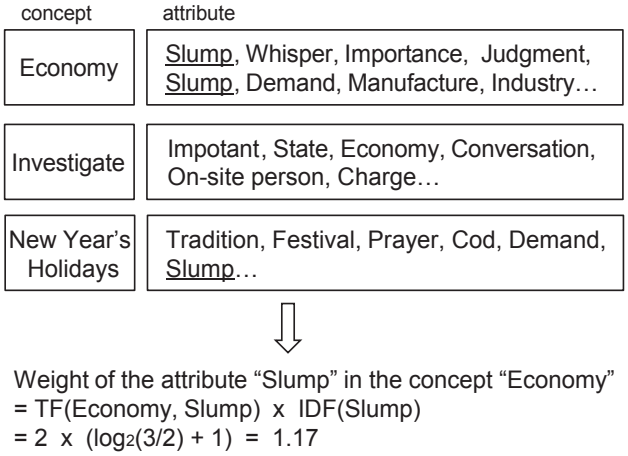

Figure 3: Specific example of weighting

This example presupposes that the total number of concepts is three. $TF(economy, slump)$ becomes two, because the attribute "slump" appears twice in the concept of "economy". $IDF(slump)$ becomes $log_2(3/2) + 1 = 0.585$,

because the number of concepts that have "slump" for an attribute is two and the total number of concepts is three.

4.4 Concept deletion via IDF threshold setting

Concepts can be deleted from the *IDF* threshold calculated in Section 4.3. by adjusting the IDF threshold setting. If the *IDF* for a concept is set too small, this concept will appear as an attribute in numerous other concepts. If this occurs, the concept appears in a vast number of co-occurrence ranges, and loses importance when defining other concepts. Examples of such concepts include the English words "the", "is", and "this". (Note that these examples are different from the Japanese words that fulfill similar roles that were eliminated when this proposed method was tested on a Japanese newspaper.) If the IDF for a concept is excessively large, it is thought that it is too high specific concept. It is thought that this process performs the CB refinement.

5 Evaluation and Validation

In this section, CB evaluation and validation methods are discussed. When evaluating a CB that has been refined by the process explained in section 4.4, the *IDF* threshold is set to several phases and the CB is evaluated for each phase.

5.1 Evaluation Method

Evaluations are carried out using an *X-BC* evaluation set. This evaluation set is composed of three concepts, X, B, and C. In Table 3, specific *X-BC* evaluation set examples can be seen. Concept B entries have some relations with Concept X entries, but Concept C entries do not.

Table 3: Specific X-BC evaluation set examples

Concept X	Concept B	Concept C
Art	Impression	Both
Situation	Consultation	Error
Write	Paper	Space
Bad Crop	Field	Keep Up
Festival	Lively	Confusion
Tea	Long-established Shop	Pail
…	…	…

This evaluation method calculates *DoA(X,B)* and *DoA(X,C)*, after which the *DoA* values are compared. If a CB is built correctly, *DoA(X,B)* should have a value that is bigger than *DoA(X,C)* because human beings can detect a relationship between concept X and concept B. Therefore, answers are considered correct when *DoA(X,B)* is bigger than *DoA(X,C)*. In this study, the total number of evaluation sets made by plural people is 500.

5.2 Evaluation and Validation at each threshold

First, an evaluation is carried out on the result of the IDF upper limit threshold. This evaluation deletes any concepts with IDF values larger than the threshold setting. Table 4 shows the result of this evaluation. In addition, it should be noted that the correct answer rate is calculated by using only the sets that contains all three (X, B and C) concepts while remaining within the threshold. The number of existing sets expresses the number of sets that contain all the (X, B and C) concepts. A set ration is a ration of number of the left sets.

Table 4: Evaluation result with IDF upper limit threshold

Threshold	Correct answer rate (%)	Number of existing sets	Set ration
14	**59.6**	319	63.8
13	59.4	318	63.6
12	59.4	310	62.0
11	59.2	282	56.4
10	56.2	217	43.4
9	42.4	118	23.6
8	29.1	55	11.0
7	22.2	18	3.6

In the case of an upper limit threshold of 14, there were no deleted concepts. Thus, the correct answer rate for the CB itself (for one newspaper year) was 59.6%. In addition, since the correct answer rate for all other threshold was less than 59.6%, no refinement effect was seen by application of upper limit threshold concept deletion. Next, an evaluation was carried out for the *IDF* lower limit threshold result. This evaluation deletes *IDF* concept values that are lower than the minimum threshold. Table 5 shows these evaluation results.

Table 5: IDF lower limit threshold evaluation results

Threshold	Correct answer rate (%)	Number of existing sets	Set ration
1	59.6	319	63.8
2	61.8	319	63.8
3	65.4	315	63.0
4	70.5	305	61.0
5	78.4	241	48.2
6	79.7	128	25.6
7	64.3	56	11.2
8	31.6	19	3.8

The correct answer rate reached the highest level when the lower limit threshold was set to 6. However, these rates cannot be compared because the number of existing sets is different depending on threshold. Accordingly, evaluations were carried out using the 128 evaluation sets that remained

in the case of a lower limit threshold of 6. Table 6 shows the evaluation results.

Table 6: Evaluation result with 128 evaluation sets

Threshold	Correct answer rate (%)
TH	58.6
1	69.5
2	68.8
3	72.7
4	78.1
5	82.8
6	79.7

As a target for comparison with the correct answer rate, a TH that resembles the degree calculation technique based on the distance on the thesaurus is utilized[5]. In the case of a lower limit threshold of 5, the evaluation result is 82.8%, which is the highest correct answer rate. It should be noted that all correct answer rates are higher than TH.

Next, the IDF value distribution was validated. To accomplish this, the distribution of all concepts (316,319) is first investigated. Table 7 shows the IDF distribution of all concepts.

Table 7: IDF distribution (All concepts)

IDF level section	Number of concepts	Accumulation rate (%)
0-1	0	0.00
1-2	2	0.00
2-3	33	0.01
3-4	85	0.04
4-5	542	0.21
5-6	1832	0.79
6-7	4634	2.25
7-8	11248	5.81
8-9	24081	13.42
9-10	51335	29.65
10-11	126690	69.70
11-12	83553	96.12
12-13	12010	99.91
13-14	274	100.0

In Table 7, the IDF level column contains "greater than left value, less than right value". For example, row "1-2" means that the value is greater than 1 but less than 2 of the IDF. In addition, the distribution of concepts in the evaluation sets (Concepts X, B and C) is investigated. Table 8 shows the IDF distribution of these evaluation sets.

Most concepts belonging to the IDF column are greater than 11 and smaller than 10. On the other hand, the peaks of the three concepts (X, B and C) exist from IDF level 6 to 9. From this, it can be seen that there is difference in word association trends between newspaper usage and human beings. These results indicate that the low IDF value is a word (concept) that appears in numerous articles. From Table 7 and Table 8, concepts that a human being can associate (concepts in evaluation sets) appear with low IDF values in the CB. This result may be due to the fact that human beings use words in conversational contexts that would be unclear and difficult to understand in newspaper articles. All CB evaluation results exceed the distance provided by a thesaurus, so it can be said that our proposed CB creation functions well. The tendency for a gap to be included in the provided concepts may dissolve when using a source of information that more closely conforms to the CB usage purpose.

Table 8: IDF distribution (Evaluation sets)

IDF level section	Number of Concept X	Number of Concept B	Number of Concept C
0-1	0	0	0
1-2	0	0	0
2-3	0	3	2
3-4	0	12	3
4-5	4	67	17
5-6	34	129	56
6-7	49	120	62
7-8	86	74	58
8-9	76	44	78
9-10	82	21	68
10-11	44	11	48
11-12	14	7	18
12-13	5	0	3
13-14	0	0	1

6 Conclusion

This paper describes a method for creating CBs easily and automatically from document groups such as newspaper articles, academic papers, and Web articles that have not been analyzed. In a CB, the meanings of various natural language phrases (called concepts) are associated with other phrase sets (called attributes) in order to detect indirect relationships. In this paper, one sentence (a range divided by periods within a newspaper article) was used to define a co-occurrence range, and concept and attributes pairs were acquired from within this range. In our CB evaluations using the proposed method, a correct answer rate is 82.8% was obtained, which is higher than resemblance degree calculation technique provided by a thesaurus.

It is desirable to make CBs using information sources that are applicable to the CB usage purpose because differences occur between the IDF distribution values of concepts extracted from newspaper articles and evaluation sets made from human conversation. However, it is clear that, by choosing an appropriate information source, the method proposed in this paper would permit the creation of CBs suitable to their usage purposes.

Acknowledgment

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (Young Scientists (B), 24700215).

References

[1] K. Kojima, H. Watabe, and T. Kawaoka. "A Method of a Concept-base Construction for an Association System: Deciding Attribute Weights Based on the Degree of Attribute Reliability"; Journal of Natural Language Processing, Vol.9 No.5, pp.93—110, 2002.

[2] H. Watabe and T. Kawaoka. "The Degree of Association between Concepts using the Chain of Concepts"; Proc. of SMC2001, pp.877—881, 2001.

[3] The Mainichi Newspapers. "CD- Mainichi '95 data collection", 1995.

[4] T. Tokunaga. "*Jyouhou Kensaku To Gengo Syori*"; University of Tokyo Press, 1999.

[5] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama. "*Nihongo Goi Taikei*"; *Iwanami Shoten*, 1997.

[6] M. Nagao. "*Iwanami Kouza SoftWare Kagaku 15 Sizen Gengo Syori*"; *Iwanami Syoten*, 1996.