Meaning Judgment Method for Alphabet Abbreviation Using Wikipedia and Earth Mover's Distance

Seiji Tsuchiya[1], **Misako Imono**[2], **Eriko Yoshimura**[1], **and Hirokazu Watabe**[1]
[1]Dept. of Intelligent Information Engineering and Science, Faculty of Science and Engineering,
Doshisha University, Kyo-Tanabe, Kyoto, Japan
[2]Dept. of Information and Computer Science, Graduate School of Engineering,
Doshisha University, Kyo-Tanabe, Kyoto, Japan

Abstract - *Recently, not only the person and things, but also words are imported by internationalization and informationization, and the scene using the loanword is increasing. However, these expressions are hard to understand for a child and the elderly person. Therefore, when such an expression is used for sentences, it might be hindered to understand entire sentences. Moreover, because an original word is omitted, it is likely to become the same expression as other words. Therefore, the alphabet abbreviation often has the polysemy. In this paper, a method of extracts an alphabet abbreviation from a sentence and judges the meaning of the expression is proposed. This method selects a correct meaning from two or more meanings of the alphabet abbreviation that suit for sentences, judging by using Wikipedia and Earth Mover's Distance. Moreover, a correct meaning is judged by the association mechanism, using an original knowledge base that defines the concept of the word. The accuracy of the proposed method was 74%.*

Keywords: alphabet abbreviation, Wikipedia, Concept Base, Degree of Association, Earth Mover's Distance (EMD)

1 Introduction

Recently, not only the person and things, but also words are imported by internationalization and informationization, and the scene using the loanword is increasing. The loanword is often used as the alphabet expression and the katakana expression, etc. in Japan. However, these expressions are hard to understand for a child and the elderly person. Therefore, when such an expression is used for sentences, it might be hindered to understand entire sentences. Especially, the alphabet abbreviation is the classic example. For instance, there is an expression by "IC". It is used when the word is omitted, and composed by initial of a certain word. If original word of the alphabet abbreviation is not understood, the expression is not understood though such an alphabet abbreviation is convenient. Moreover, because an original word is omitted, it is likely to become the same expression as other words. Therefore, the alphabet abbreviation often has the polysemy. Previous example, "IC" has two or more meanings such as "Integrated circuit" and "Interchange in the expressway", etc.

In this paper, a method of extracts an alphabet abbreviation from a sentence and judges the meaning of the expression is proposed. This method selects a correct meaning from two or more meanings of the alphabet abbreviation that suit for sentences, judging by using Wikipedia and Earth Mover's Distance. Moreover, a correct meaning is judged by the association mechanism, using an original knowledge base that defines the concept of the word.

2 Proposed Method and Elemental Technique

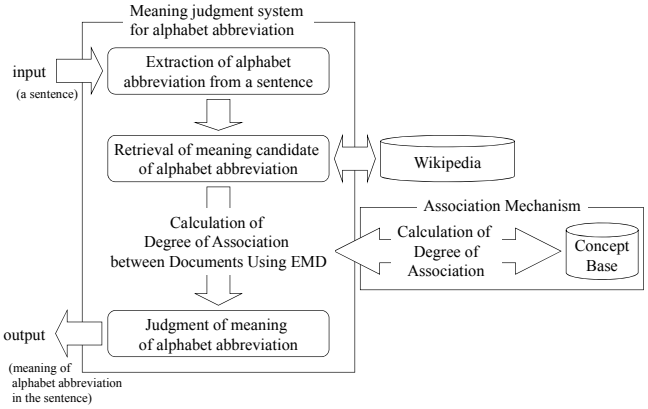

Fig. 1 Outline of the proposed meaning judgment method for alphabet abbreviation

Figure 1 shows the outline of the proposed meaning judgment method for alphabet abbreviation. When a sentence is inputted an alphabet abbreviation is extract which included in the sentence. The extracted alphabet abbreviation is retrieved with Wikipedia [1], so the original word of the alphabet abbreviation is obtained. The original word of the alphabet abbreviation judges the word that suit for sentences by evaluating the relativity of input sentences and Wikipedia's explanation sentences of those words in two or more cases. The relations between the sentences are evaluated by technique of Earth Mover's Distance. And, the relations

between the words are evaluated by the idea of the Concept Base [2][3] that defines the concept of the word and the idea of the Degree of Association [4][5] that calculates the association between words. A lot of words like the proper noun etc. are not existed in the Concept Base because the Concept Base is constructed by using the national language dictionary. However, it is necessary to conceptualize the word to calculate the Degree of Association, so the word that doesn't exist in the Concept Base is automatically conceptualized by using information on Web.

2.1 Concept Base

A Concept Base is a large-scale database that is constructed both manually and automatically using words from multiple electronic dictionaries as concepts and independent words in the explanations under the entry words as concept attributes. In the present research, a Concept Base containing approximately 90,000 concepts was used, in which auto-refining processing was carried out after the base had been manually constructed. In this processing, attributes considered inappropriate from the standpoint of human sensibility were deleted and necessary attributes were added.

In the Concept Base, concept A is expressed by attributes a_i indicating the features and meaning of the concept in relation to a weight w_i denoting how important an attribute a_i is in expressing the meaning of concept A. Assuming that the number of attributes of concept A is N, concept A is expressed as indicated below. Here, the attributes a_i are called primary attributes of concept A.

$$A = \{(a_1, w_1), (a_2, w_2), ..., (a_N, w_N)\} \qquad (1)$$

Because the primary attributes a_i of concept A are taken as the concepts defined in the Concept Base, attributes can be similarly elucidated from a_i. The attributes a_{ij} of a_i are called the secondary attributes of concept A. Figure 2 shows the elements of the concept "train" expanded as far as the secondary attributes.

train	train, 0.36	locomotive, 0.21	railroad, 0.10	...	a_i, w_i	Primary Attributes
	train, 0.36	locomotive, 0.21	railroad, 0.10	...	a_{i1}, w_{i1}	Secondary Attributes
	locomotive, 0.21	streetcar, 0.23	subway, 0.25	...	a_{i2}, w_{i2}	
	:	:	:	:	:	
	a_{1j}, w_{1j}	a_{2j}, w_{2j}	a_{3j}, w_{3j}	...	a_{ij}, w_{ij}	

Concept

Fig. 2 Example demonstrating the concept "train" expanded as far as secondary attributes

2.2 Methods of Automatically Expanding the Concept Base

If terms that are not in the Concept Base (undefined terms) are not given attributes, it will not be possible to seek the Degree of Association between undefined terms and other terms. For this reason, we propose a method of conceptualizing undefined terms based on Web data, which currently the largest collection of linguistic data, and then adding these to the Concept Base.

2.2.1 Making Concepts of Undefined Terms

In order to conceptualize an undefined term, the attributes and weighting of the term are acquired from the Web using the procedure described below:

(1) A search engine is used to search for the entered but undefined term as a key word and obtain the content of the "top 100 search results" page.
(2) A morphological analysis is then applied to the document collection, and unnecessary data, such as HTML tags, will be removed and independent terms extracted.
(3) From among the independent terms obtained, only those that exist in the Concept Base are extracted as the attributes of undefined terms.
(4) The attribute frequency is multiplied by the SWeb-idf value, a statistically investigated idf of terms on the Web, and the value obtained is set as the attribute weighting. These are rearranged in order of the weighting. SWeb-idf will be explained in the following section. Attributes that do not exist in the SWeb-idf database are considered as terms that do not exist much on the Web, so they were multiplied using the maximum value of SWeb-idf.

2.2.2 SWeb-idf

SWeb-idf (Statics Web-Inverse document frequency) is an idf value that statistically examines the idf of terms on the Web. First, it generates 1,000 proper nouns that are randomly chosen. A search was carried out for each of the 1,000 terms created and the content of the top ten search result pages was obtained for each single term. As a consequence, the number of search result pages amounted to 10,000. Because we were able to obtain a number of terms that amounted to about the same 90,000 terms contained in the Concept Base from these 10,000 pages, which is a knowledge base that extracted the concepts (terms) from sources such as multiple Japanese language dictionaries and newspapers, we considered the 10,000 pages to be the information space for all data on the Web. SWeb-idf, which expresses the idf value of the terms within those pages, can be found using equation 2 below.

$$SWeb - idf(t) = \log\frac{N}{df(t)}, \quad (N=10000) \quad (2)$$

The terms and idf values obtained from this were registered in the database. The $df(t)$ part of the equation is the number of concept t pages that appear within all of the document spaces (10,000 pages).

2.2.3 Weighting Method Using Frequency of Appearance within the Attribute

Although the weighting of an attribute of an undefined term can be found using SWeb-idf, a distortion of the Concept Base frequency data will occur if the Web data weighting is used as is and added to the Concept Base. This is because the frequency data for terms differ in the Web data and the Concept Base and their weighting values change. As a result, it is only used when SWeb-idf obtains an undefined term's attribute candidate and not used for the attribute a weighting of the undefined term. Thus, when an undefined term is added to the Concept Base the frequency data of the Concept Base must be used to assign a weighting. This is why we propose a weighting method that takes the Concept Base attribute space into consideration as a means of assigning a weighting to the attribute of an undefined term. Since the attribute assigned to a concept is a term that expresses characteristics, it can be understood as being the explanatory text of that concept. The frequency of appearance of an attribute in this document space is considered the probability of the attribute relative to the concept.

It is possible to see the n order attribute space for the concept as a set of explanatory text for the concept. The frequency of appearance calculated from this n order attribute is called the frequency of appearance within the n order attribute. In this paper, the secondary attribute space is used. Based on the thinking behind the weighting of the tf-idf, if the frequency of appearance of the secondary attribute of undefined term attribute A is $freq(A)$, the total number of undefined term primary attributes is R, the idf value of the Concept Base space for undefined terms is $cidf(A)$, then the weighting $wc(A)$ can be expressed as shown in the following equation:

$$wc(A) = \frac{\log(freq(A))}{\log(R)} cidf(A) \qquad (3)$$

2.3 Calculating of the Degree of Association

For concepts A and B with primary attributes a_i and b_i and weights u_i and v_j, if the numbers of attributes are L and M, respectively ($L \leq M$), the concepts can be expressed as follows:

$$A = \{(a_1, u_1), (a_2, u_2), ..., (a_L, u_L)\} \qquad (4)$$

$$B = \{(b_1, v_1), (b_2, v_2), ..., (b_M, b_M)\} \qquad (5)$$

The Degree of Identity $I(A, B)$ between concepts A and B is defined as follows (the sum of the weights of the various concepts is normalized to 1):

$$I(A,B) = \sum_{a_i=b_i} \min(u_i, v_j) \qquad (6)$$

The Degree of Association is calculated by calculating the Degree of Identity for all of the targeted primary attribute combinations and then determining the correspondence between primary attributes. Specifically, priority is given to determining the correspondence between matching primary attributes. For primary attributes that do not match, the correspondence between primary attributes is determined so as to maximize the total degree of matching. Using the degree of matching, it is possible to give consideration to the Degree of Association even for primary attributes that do not match perfectly.

When the correspondences are thus determined, the Degree of Association $R(A, B)$ between concepts A and B is as follows:

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

In other words, the Degree of Association is proportional to the Degree of Identity of the corresponding primary attributes, and the average of the weights of those attributes and the weight ratios.

2.4 Degree of Association between Documents Using EMD

When seeking the degree of similarity between a search request and a search target, no matter how accurately the relevance between terms can be defined, if the calculation cannot take place based on the values, it will be impossible to find the precise degree of similarity between the documents. A variety of methods can be used for the calculation. For instance, one method would be to perform the calculation by correlating the terms in order beginning from the highest degree of relevance between the terms. A method that involves a one-to-one correlation can only correlate to the smallest number of terms between the search request and the search target. For example, if the search request has three terms and the search target has 100 terms, 97 of the search target terms will not be subjected to calculation. Furthermore, it is believed that when performing the actual search, users will not enter many terms in the search request, so the assumption is that there will be a large difference in the number of terms in the search request and search target. Therefore, it is necessary to consider the importance of terms in the text and the relevance between them and to be flexible in handling M relative to N.

For this reason, the Earth Mover's Distance (EMD) [6], which has been drawing attention in the field of similar imagery searching, has been employed in this study as a method that calculates the degree of similarity between documents. The EMD is an algorithm that seeks the optimal solution for transportation costs in a transportation problem. As a result, if the weighting between the demand point and the supply point and the distance between these points are

defined, it can be used to solve any type of problem. By employing the EMD and taking the weighting of terms and the relevance between terms into consideration, correlation can be flexible and the degree of similarity between sentences can be found.

2.4.1 What is the EMD

The EMD is a distance scale that calculates by means of the Hitchcock transportation problem, which is one type of linear programming problem. Given two discrete distributions, it is defined as the minimum cost of converting one distribution to the other distribution. The transportation problem is the problem of solving transportation from the supply point to the demand point in order to satisfy the demand at the demand point at minimum cost.

When seeking the EMD, the two distributions are expressed as sets that have been assigned element weightings. If one of the distributions P is expressed as a set, the expression becomes, $P = \{(p_1, w_{p1}), \ldots, (p_m, w_{pm})\}$. Distribution P is currently expressed as having m number of characteristics. p_i represents the characteristics, while w_{pi} represents the weighting of the characteristics. In like manner, if the other distribution Q is expressed as a set, the expression becomes, $Q = \{(q_1, w_{q1}), \ldots, (q_n, w_{qn})\}$. As for the EMD calculation, even if the number of characteristics for both distributions differs, it has a characteristic that allows the calculation to take place. Let us assume that the distance between p_i and q_j is d_{ij} and the distance between all features is $D = [d_{ij}]$. If we assume the amount of transportation from p_i to q_j to be f_{ij}, the total amount of transportation becomes $F = [f_{ij}]$. Here, we will find the amount of transportation F, which creates the minimum cost function shown in equation 7, and calculate the EMD.

$$WORK(P,Q,F) = \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} f_{ij} \qquad (7)$$

However, when minimizing the above cost function, the following restrictions must be satisfied.

$$f_{ij} \geq 0,\ 1 \leq i \leq m,\ 1 \leq j \leq n \qquad (8)$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i},\ 1 \leq i \leq m \qquad (9)$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j},\ 1 \leq j \leq n \qquad (10)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right) \qquad (11)$$

In this case, we know that the amount of transportation in equation 8 is positive and we also know that transportation goes one way, from p_i to q_j. Equation 9 indicates that transportation cannot take place above the weighting of the transportation source p_i. Equation 10 indicates that acceptance cannot take place above the weighting of the transportation destination q_j. Finally, equation 11 indicates the upper limit of the total amount of transportation and is limited by the smaller of the sum total of either the transportation destination or transportation source. The EMD between distributions P and Q can be found as indicated below by using the optimal total amount of transportation F found under the limitations indicated above.

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \qquad (12)$$

The reason the optimal cost function $WORK(P, Q, F)$ is used as is here as the EMD is that the cost function depends on the sum total of the weighting of either the transportation source or the transportation destination. So, that influence will be eliminated by normalization.

2.4.2 Applying EMD to Document Search

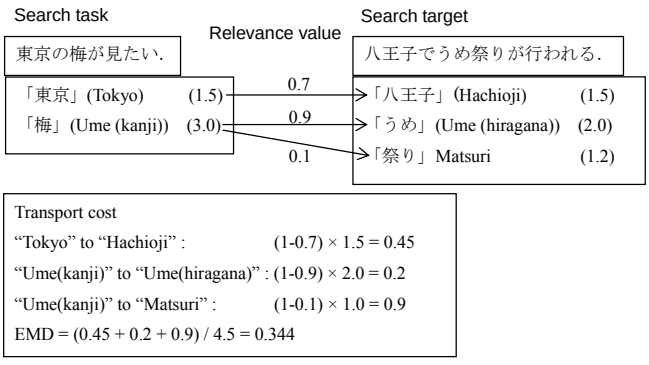

Fig. 3 Examples of Applying EMD to Document Search

Figure 3 shows examples of applying EMD to document search. To apply EMD to document search, the demand point and supply point, demand volume and supply volume, and the distance between each demand point and supply point must be defined. For the demand point, the index term for the search task is assigned, while for the supply point, the index term for the search result is assigned. The demand volume and supply volume each use the tf-idf weighting [7], which concerns index terms. The distance between the demand point

and supply point can be considered to be the relevance between index terms and, thus, can be found in the proposed methodology by the Degree of Association that uses the Concept Base. Since the value of the Degree of Association will be larger as the relevance increases, it will be converted into a value in which the Degree of Association value will be subtracted from 1. The calculation of EMD is located at the bottom of Figure 3. The reason that the amount of transportation between "ume" and "matsuri" is 1 is because a weighting of 2 was transported from "ume (kanji)" to "ume (hiragana)" and the excess weighting of "ume (kanji)," 1, was transported to "matsuri. " The weighting is transported in this manner to terms with a high degree of relevance and the transportation will take place until the supply volume disappears or the demand volume is satisfied. In this way, a flexible M versus N that considers relevance and weighting between index terms is possible. As a characteristic of the EMD, if the value of the distance between index terms is from 0 to 1, then EMD also becomes a value from 0 to 1. Additionally, if there is similarity between documents, the value falls, and if there is a lack of similarity, the value rises. Thus, document retrieval is realized by presenting documents to the user in sequence beginning with documents with low values.

3 Meaning Judgment Method of Alphabet Abbrebiation

The proposed meaning judgment method is composed of three processing as shown in Figure 1: extraction of alphabet abbreviation from a sentence, retrieval of candidate meaning of alphabet abbreviation and judgment of meaning of alphabet abbreviation.

3.1 Extraction of alphabet abbreviation from a sentence

The rule of extracting the alphabet abbreviation from sentence is three of the following:

(1) The string of alphabet character composed of the capital letter and the small letter of one character or more, and it is extracted from sentence.

(2) The string of character that there is a figure after the alphabet character string is extracted. For instance, correspond to "CO2" etc.

(3) The string of character which including the sign is not extracting to the string of alphabet character. For instance, correspond to "http://www" etc.

3.2 Retrieval of meaning candidate of alphabet abbreviation

The alphabet abbreviation extracted by the rules of section 3.1 is retrieving with Wikipedia, and the original word is obtained from candidate of the word (meaning).

When the candidate meaning is one, the word and the detailed explanation that becomes the original word of the alphabet abbreviation are described in Wikipedia. Accordingly, the word that becomes the original word is a meaning of the alphabet abbreviation in this case.

On the other hand, the words that become two or more original words are enumerated in Wikipedia when there are two or more meaning candidates. Therefore, all enumerated words are acquired as a meaning candidate in this case.

3.3 Judgment of meaning of alphabet abbreviation

When two or more meaning candidates are acquired by processing of section 3.2, it is necessary to judge a correct meaning from several meaning candidate. A correct meaning is judged from the evaluation of meaning association between an input sentence and the meaning candidates. The meaning candidates are expressed as explanation sentences in Wikipedia. The Concept Base and the method of calculation of Degree of Association explains in Chapter 2 are used for the evaluation of meaning association.

The words included in the explanation sentences of the meaning candidates are conceptualized by using the attributes which exist and defined in the Concept Base. It is explained in section 2.1. When the words included in the explanation sentences of the meaning candidates do not exist in the Concept Base, they are conceptualized by the explained method in section 2.2. Moreover, the words included in the input sentence are similarly conceptualized. In addition, the weights of words are given by tf-idf technique [7].

The association between the words included in the explanation sentences of the meaning candidates and the words included in the input sentence are calculated by the method of the calculating of the Degree of Association explained in section 2.3. And, the association between the explanation sentences of the meaning candidates and the input sentence is calculated by the method of the EMD explained in section 2.4 using results of the association between words. As a result, a correct meaning of the alphabet abbreviation is decided by strongest relation between the explanation sentences of the meaning candidate and the input sentence.

4 Performance Evaluation of the Proposed Meaning Judgment Method for Alphabet Abbreviation

100 newspaper articles that contained the string of alphabet character were selected at random and used for evaluation data. A correct meaning of the alphabet abbreviation was judged by three subjects. When the processing explained by section 3 was done to the 100 newspaper articles, the alphabet abbreviations of 54 words with two or more meanings have been extracted.

Figure 4 shows the comparative results of proposed method using EMD and method no using EMD. The accuracy of the proposed method was 74%. The accuracy of proposed method using EMD was improved 12% to compare with the method no using EMD. By these results, the proposed method is able to say effective.

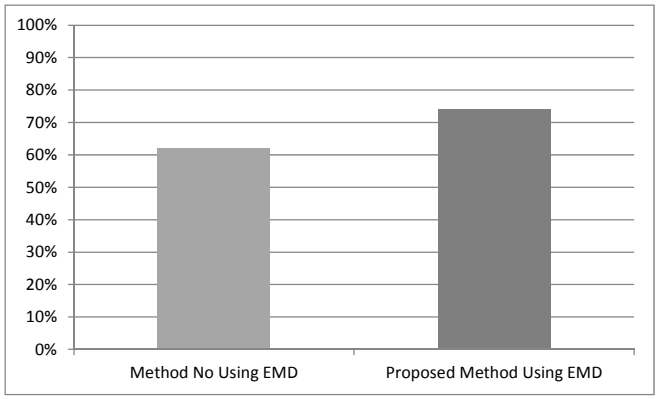

Fig. 4 Result of the proposed meaning judgment method for alphabet abbreviation

5 Conclusions

A method which extracts an alphabet abbreviation from a sentence and judges the meaning of the expression was proposed in this paper. This method selects a correct meaning from two or more meanings of the alphabet abbreviation that suit for sentences, judging by using Wikipedia and Earth Mover's Distance (EMD). Moreover, a correct meaning was judged by the association mechanism using an original knowledge base that defined the concept of the word.

100 newspaper articles that contained the string of alphabet character were selected at random and used for evaluation data. The alphabet abbreviation of 54 words was extracted from 100 newspaper articles. A correct meaning of the alphabet abbreviation was judged by three subjects, and the accuracy of this method was 74%. By the result, we believe that the proposed method is effective.

Acknowledgements

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] Wikipedia, http://ja.wikipedia.org/wiki

[2] Kojima, K., Watabe, H. and Kawaoka, T.: A Method of a Concept-base Construction for an Association System: Deciding AttributeWeights Based on the Degree of Attribute Reliability. Journal of Natural Language Processing. 9(5), pp.93–110, 2002.

[3] N. Okumura, E. Yoshimura, H. Watabe, and T. Kawaoka, "An Association Method Using Concept-Base", KES 2007/WIRN2007, Part I, LNAI4692, pp.604–611, 2007.

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

[5] Hirose, T., Watabe, H. and Kawaoka, T.: Automatic Refinement Method of Concept-base Considering the Rule between Concepts and Frequency of Appearance as an Attribute. Technical Report of the Institute of Electronics, Information and Communication Engineers. NLC2001-93, pp.109–116, 2002.

[6] Y. Rubner, C. Tomasi, and L. Guibas: "The earth mover's distance as a metric for image retrieval", Int. J. Comput. Vision, Vol. 40, pp. 99–121, 2000.

[7] Salton, G. and Buckley, C. , "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, Vol.41, No.4, pp.513-523, 1988.