

Meaning Estimation Method of Alphabetical Abbreviation Using Lexical Conceptualization and Wikipedia

Kazuto Goto, Seiji Tsuchiya, and Hirokazu Watabe

Abstract—Alphabetical abbreviations are often used; they are composed of initials of each English word. However, alphabetical abbreviations have problem with the polysemy. In this paper, we propose the method to estimate the meaning of an alphabetical abbreviation. The proposal method regards meaning estimation of an alphabetical abbreviation as meaning estimation of unknown word. This method uses the concept base, and, Calculation of Degree of Association or Earth Mover's Distance. They allow conceptualization of a word and evaluation of semasiological association between conceptualized words. And, Wikipedia is used to complement lack of information due to an alphabetical abbreviation. This Paper used 129 articles to evaluate the proposal method. The experiments showed that the accuracy of the proposal method was 79% and this method was effective than other method.

Index Terms—Alphabetical abbreviation, meaning estimation, conceptualization, Wikipedia.

I. INTRODUCTION

Loanwords are widely used in Japan, and, many loanwords are often written in katakana; katakana is one of Japanese notation. Recently, it is likely to tend to increase using loanwords, and, it is found that loanword is not written in katakana. Especially in English, alphabetical abbreviations are often used; alphabetical abbreviations are composed of initials of each English word. For example, “IC” is an alphabetical abbreviation. However, alphabetical abbreviations have problem with the polysemy. For instance, “IC” has two or more meanings such as “Integrated Circuit” and “InterChange in the express way”. Due to the polysemy, some people cannot understand easily and precisely the information which is included the alphabet abbreviation.

II. PREVIOUS RESEARCH

Based on above, this paper proposes the method to estimate meaning of alphabetical abbreviation which is used in Japanese. As similar researches, the research of [1] is reported. To achieve automatic estimation of meaning of alphabetical abbreviation, this research focuses attention on an expression; in sentences such as newspaper, meaning of alphabetical abbreviation is noted in brackets. However, such expressions are not applied to all alphabetical abbreviations.

Manuscript received December 30, 2016; revised April 3, 2017. This work was supported in part by JSPS KAKENHI Grant Number 16K00311.

Kazuto Goto is with the Graduate School of Science and Engineering, Doshisha University, Kyotanabe, Japan (e-mail: eup1102@mail4.doshisha.ac.jp).

Seiji Tsuchiya and Hirokazu Watabe are with Faculty of Science and Engineering, Doshisha University, Kyotanabe, Japan (e-mail: stsuchiya@mail.doshisha.ac.jp, hwatabe@mail.doshisha.ac.jp).

[2] looks at biomedical field in which alphabetical abbreviation occurs frequently. This research estimates meaning of alphabetical abbreviation by using corpus; the corpus is constructed based on the dataset which is extracted from database about medicine. However, because this method emphasizes the specific field of medicine, it is difficult to apply generally.

Also, although alphabetical abbreviation is not targeted, method to resolve word sense disambiguation is researched. [3] achieves word sense disambiguation by using hyperlink which is reference information of each article in Wikipedia when inputting a general word and a text including the word. This method has a problem that processing takes time and effort, since the method requires human hand for mapping to WordNet which is concept dictionary after labeling the input word by using Wikipedia. Moreover, as entity disambiguation which resolve ambiguity by mapping a word to entry of knowledge base, [4] constructs neural network based on information which is collected from Wikipedia and evaluates similarity between a input word and each entry. However, because this method needs to treat set with “Context, Mention, and Entity,” convenience of operation is low.

In this paper, we propose the method to estimate meaning of alphabetical abbreviation which has polysemy. This method estimates meaning of alphabetical abbreviation by combining the mechanism which can evaluate semasiological distance of every word. Concretely speaking, we use concept base [5] and Wikipedia [6]; concept base allows lexical conceptualization which can associate a concept with various concepts, and Wikipedia is described as the dictionary which has most words in the world. In addition, to evaluate semasiological distance of a word which is conceptualized, we use Calculation of Degree of Association [7] or calculation method of degree of association between sentences which is based on Earth Mover's Distance [8]. The proposal method resolves the polysemy of alphabetical abbreviation by using these mechanisms, and estimates meaning of alphabetical abbreviation.

III. CONCEPT OF THE PROPOSAL METHOD

Fig. 1 shows the schematic view of the meaning estimation method of alphabetical abbreviation. First, an alphabetical abbreviation is extracted from an input sentence. Next, this method searches for the alphabetical abbreviation on Wikipedia. If a meaning of the alphabetical abbreviation is one, the meaning is output. Also, if meanings are plural, semasiological distance between each meaning and the input sentence is evaluated, and most kindred meaning is determined. In this case, lexical conceptualization is

conducted. The above-mentioned “meaning” shows a word which expresses meaning of alphabetical abbreviation, that is, Japanese expression of English word which becomes source of alphabetical abbreviation is defined as ”meaning”. For example, if the proposal method estimates meaning of previous described alphabetical abbreviation “IC,” “Integrated Circuit” or “InterChange” is output as “meaning”.

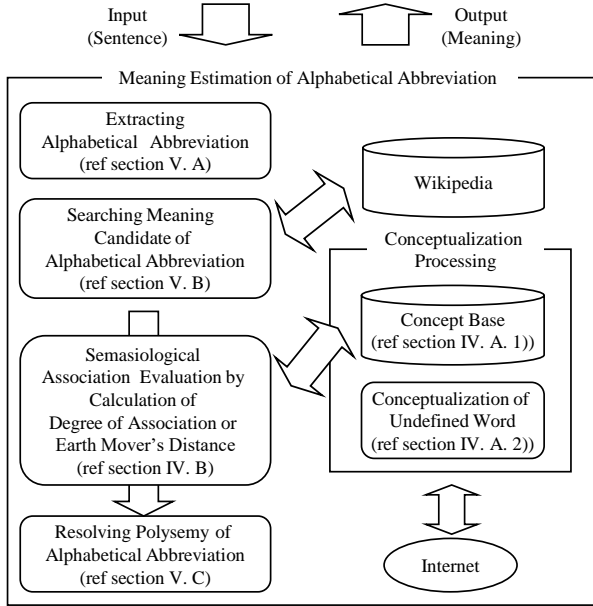


Fig. 1. The schematic view of the proposal method.

IV. ELEMENT TECHNOLOGY

A. Lexical Conceptualization

1) Concept base

The concept base [5] is a large-scale database which is consisted of concepts and attributes; concepts are direction word in such as multiple computerized dictionaries, and attributes are independent words which are used in the word meaning sentence. The concept base has about 90,000 concepts. In the concept base, a concept A is consisted of m attributes a_i and weight w_i which shows importance of the attribute. These are given as follows. The attribute a_i is called the primary attribute of the concept A .

$$A = \{(a_1, w_1), (a_2, w_2), \dots, (a_m, w_m)\}$$

As quite characteristic of the concept base, the word which is the attribute is inevitably defined as a concept. So, the attribute a_i which is the primary attribute of concept A is considered as a concept, and it is possible to derive additional attributes. The attribute a_{ij} which is derived from the concept a_i is called the secondary attribute of the original concept A . We can define semantic character of a concept by the concept base, and evaluate semasiological distance of words by using chained design. The detail is shown in section IV.B.

Table I shows the example of the concept base. Although the actual concept base is constructed by Japanese, Table I is written in English due to readability.

2) Conceptualization of undefined word

The concept base has the advantage that it is large scale

and high-quality. However, the concept base is disadvantageous in that it cannot cover all words. Therefore, the undefined word which is not registered in the concept base is not conceptualized, and we cannot evaluate semasiological distance.

So, we deal with the undefined word by using linguistic information on Web [9]. Concretely speaking, the undefined word is searched for information on Web search engine [10], and the contents of search result page are obtained from the top 100 search result. Independent words which are registered in the concept base are extracted from the contents, and these words become the primary attribute of the undefined word. Also, weight of the primary attribute is calculated by applying $tf \cdot idf$ which is widely used in information search field [11].

tf value which is completeness of word is calculated by dividing $tfreq(Word_A, A)$ by $tnum(A)$; $tfreq(Word_A, A)$ is appearance frequency of independent word $Word_A$ in search result page A , and $tnum(A)$ is the total number of independent words in search result page A . Formula for computation is below.

$$tf(Word_A, A) = tfreq(Word_A, A) / tnum(A)$$

Next, idf value which is specificity of word is calculated from Statics Web Inverse Document Frequency ($SWeb_idf$) value [9]. Information of intended all document space is needed for calculation of idf value. However, it is impossible to calculate accurate idf value on Web in practice, because all information on Web is needed. So, 1,000 proper nouns which are selected at random are set as keyword respectively and are searched on Web search engine [10]. Subsequently, the contents of search result page are obtained from the top 10 search result. The collection of all independent words which is included in obtained page is considered as pseudo all information space of Web, and $SWeb_idf$ value is calculated. Formula for computation is below; N is the total number of search result pages which are obtained at the time of searching 1,000 proper nouns ($N = 10,000$), and $df(Word_A)$ is the number of search result pages which include $Word_A$.

$$SWeb_idf(Word_A) = \log\{N / df(Word_A)\}$$

Because over 90,000 words are obtained from these 10,000 pages and the concept base has about 90,000 words, these pages are considered as all information space of Web. In addition, if choice of proper nouns is changed, it is reported that $SWeb_idf$ value does not make a big change [9]. As stated above, weight w which is assigned to the independent word $Word_A$ is defined by the following formula.

$$w = tf(Word_A, A) \cdot SWeb_idf(Word_A)$$

TABLE I: THE EXAMPLE OF THE CONCEPT BASE

Concept	Attribute
Doctor	(Physician, 0.34), (Medical Ward, 0.08), ...
Medical Ward	(Clinic, 0.25), (Operation, 0.18), ...
Patient	(Hospital, 0.52), (Nursing, 0.21), ...
...	...

Due to using linguistic information on Web, quality is lower than the concept base. However, this method has the

greater advantage that we can conceptualize all words. Also, it is considered that enough performance for the processing in this paper is ensured according to study results of [9].

B. Evaluation Approach for Semasiological Association

1) Calculation of degree of association

Calculation of Degree of Association (DoA) is calculation to evaluate quantitatively association between one concept and another. This paper uses DoA with weight ratio [7].

Regarding arbitrary concept A and B , each primary attribute is a_i and b_j , corresponding weight is u_i and v_j . Also, the number of attributes of the concept A and B is L and M ($L \leq M$). In addition, each weight of primary attribute of concept is normalized so that summation is 1.0. In this case, Degree of Match (DoM) with weight ratio between the concept A and B is defined by the following.

$$A = (a_i, u_i) \mid i = 1 \text{ to } L, \quad B = (b_j, v_j) \mid j = 1 \text{ to } M$$

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

Next, the concept which has a low number of attributes than the other is A , and the attributes of the concept A are criteria. And, the attributes of the concept B are rearranged so that the sum of $DoM(a_i, b_{x_i})$ is max; $DoM(a_i, b_{x_i})$ is DoM with weight ratio between the attributes of concept A and B .

$$A = \{(a_1, u_1), \dots, (a_i, u_i), \dots, (a_L, u_L)\}$$

$$B = \{(b_{x_1}, v_{x_1}), \dots, (b_{x_i}, v_{x_i}), \dots, (b_{x_L}, v_{x_L})\}$$

Regarding the primary attributes of both concepts, corresponding pair is decided by the above procedure. Combination number of corresponding attributes is T .

DoA with weight ratio is calculated from the maximum value of sum of DoM with weight ratio; DoM with weight ratio is calculated from between each attribute of comparing concepts. $DoA(A, B)$ of the concept A and B is defined by the following formula.

$$DoA(A, B) = \sum_{i=1}^T \{DoM(a_i, b_{x_i}) \cdot (u_i + v_{x_i}) \cdot (\min(u_i, v_{x_i}) / \max(u_i, v_{x_i})) / 2\}$$

Hereafter, DoM with weight ratio is abbreviated to DoM, and DoA with weight ratio is abbreviated to DoA. This paper uses this degree of association [7]. DoA expresses association between one concept and another by continuous value which is 0-1.

2) Earth mover's distance

Calculation of DoA is performed by corresponding to attributes in decreasing order of association, that is to say, this calculation is applied to one-on-one attributes. So, concerning both concepts, this calculation uses only partial attributes. In this paper, to tackle the circumstances which differ in the number of attributes of both concepts, in addition to DoA, we use Earth Mover's Distance (EMD); EMD can evaluate semasiological association and can correspond to M-on-N.

EMD is distance scale which is calculated by Hitchcock transportation problem [12]. Evaluation approach for

semasiological association by using EMD applies EMD to evaluate association between concepts. EMD can quantitatively express association between concepts, and the availability of EMD is reported by study of [8].

When there are two discrete distribution, EMD means minimum cost to translate one distribution into the other. Discrete distribution expresses the set of element and weight. In calculation cost, supply quantity is weight which element of discrete distribution before translation has, demand quantity is weight which element of discrete distribution after translation has. Also, weight is transported according to distance between elements, supply quantity, and demand quantity. EMD is the pathway which can transport weight to demand center as short distance and efficiently as possible.

Fig. 2 shows schematic view which explains the semasiological association evaluation method with the use of EMD. If concept A and B are given, this method calculates the cost for translating the concept A into the concept B . Respective concepts are considered as discrete distribution of these primary attribute a_i and b_j . EMD uses the distance of between elements which configure the discrete distribution for calculation of translation cost. In this paper, this distance is considered as association between the primary attributes, and is calculated by DoM.

The distance $dis(a_i, b_j)$ between the attribute a_i and b_j is calculated by subtracting DoM from the maximum value of DoM; maximum value of DoM is 1. Also, the translating cost $cost(a_1, b_1)$ between a_1 and b_1 is calculated by multiplying the distance by weight.

Calculation of the cost is equally performed about other pair of attributes, and finally EMD is calculated by adding cost of all transportation pathways. In Fig. 2, EMD between the concept A and B is expressed as follows. The minimum of calculated EMD is evaluated by optimized calculation, and association between concepts is calculated.

$$EMD(A, B) = cost(a_1, b_1) + cost(a_2, b_2) + cost(a_2, b_3)$$

V. MEANING ESTIMATION METHOD

A. Extracting Alphabetical Abbreviation

In this paper, the meaning estimation method is aimed at alphabetical abbreviation which is composed of initials of English word. Input information is sentences which have alphabetical abbreviation. The proposal method extracts row of capital alphabets (over 2 characters) from the sentences, and the alphabets are treated as alphabetical abbreviation.

B. Searching for Meaning Candidate on Wikipedia

The proposal method does search on Wikipedia [6] for the alphabetical abbreviation which is extracted in section V.A. In consequence of the search, if a meaning candidate which explains the alphabetical abbreviation is one, the meaning candidate is output.

Also, if candidates are plural, each candidate is conceptualized to resolve the polysemy. The concept base (ref. section IV.A.1)) and conceptualization method of undefined word (ref. section IV.A.2)) are used for the conceptualization.

Meaning candidates which are extracted from Wikipedia

include information which explains such meaning, but the information becomes noise in conceptualization of the meaning. Therefore, to eliminate the noise, we apply the rule which is shown in Table II to the meaning candidate; the rule is applied beginning at the top. In addition, Wikipedia contains numerous meanings which are not alphabetical abbreviation, such as model number of a product and some sort of code. To delete such meaning, we apply the stop-word which is shown in Table III.

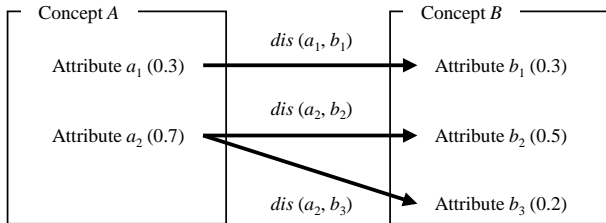


Fig. 2. The semasiological association evaluation with the use of EMD.

C. Resolving Polysemy of Alphabetical Abbreviation

If the alphabetical abbreviation has multiple meanings, it is necessary to resolve the polysemy. Concretely speaking, the proposal method evaluates semasiological distance between conceptualized meaning candidate (ref. section V.B) and input sentence including the alphabetical abbreviation. To evaluate semasiological distance, it is also necessary to conceptualize alphabetical abbreviation.

Independent words in input sentence become primary attributes of the alphabetical abbreviation, and the alphabetical abbreviation is conceptualized. The conceptualized alphabetical abbreviation is expanded to primary attributes and secondary attributes, and we can evaluate semasiological distance. In consequence, the meaning candidate which is judged to be near most semantically is the meaning of alphabetical abbreviation.

VI. EVALUATION EXPERIMENT

A. Experiment Condition

In this paper, as evaluation experiment data, alphabetical abbreviations were extracted from articles. The articles which include the alphabetical abbreviations were input sentence. The articles for the evaluation experiment were one month's national daily paper (about 12,000 articles). About 3,700 articles included row of the capital alphabets more than 2 characters. We cut articles which include stop-word (ref. TABLE III) by human hand, and 129 articles were randomly extracted. In the extracted articles, the alphabetical abbreviations which have different notation were 58. Also, we did search on Wikipedia for the 58 alphabetical abbreviations and applied the rule which is shown TABLE II and TABLE III, and consequently got 707 meanings.

In this evaluation experiment, if meaning of alphabetical abbreviation which was estimated by the proposal method corresponded to meaning in the article which included the alphabetical abbreviation, we judged output as right answer. Due to reduction, evaluation target was articles which have 1 alphabetical abbreviation.

B. Evaluation Result

This paper used (1) Degree of Association (DoA), (2) Earth Mover's Distance (EMD), (3) Vector Space Model (VSM). VSM is search model which is widely used in the field of information search, and is used for comparison of the proposal method. In this evaluation, we used VSM which employs cosine distance on distance metric and $tf \cdot idf$ on weighting. The detail of VSM is explained in [13].

TABLE II: THE RULE FOR OBTAINING THE MEANING CANDIDATE OF ALPHABETICAL ABBREVIATION FROM WIKIPEDIA

- If a meaning candidate includes stop-word (ref. Table III), the meaning candidate is eliminated.
- In a meaning candidate, the word in back of left corner bracket ([), right arrow (->), and hyphen (-) is deleted.
- In a meaning candidate, "etc." is deleted.
- In a meaning candidate, the word in front of "thing of," "abbreviation for," "ellipsis for," "popular name of," "nickname of," "appellation of," and "English abbreviated name of" is obtained.
- In a meaning candidate, the word in back of "one of," is obtained.

TABLE III: THE LIST OF STOP-WORD

model number, model code, format, series, brevity code, unit, domain, extension, denotation, symbol, coefficient, command, country code, administrative area code, prefecture code, post code, airport code, harbor code, IATA code, airline code, format code, currency code, language code, production, music, appearance, album, call letters, list, above, other

Accuracy rate of DoA is 76%, EMD is 79%, and VSM is 69%. As compared to VSM, accuracy rate of DoA is 7% high and EMD is 10% high. Also, accuracy rate of EMD is slightly higher (3%) than DoA. Also, previous research [1] which is shown section II has limiting condition for applicable alphabetical abbreviation; it is necessary to write the meaning of the alphabetical abbreviation with brackets. For the 129 articles used in this evaluation, the alphabetical abbreviations to which we could apply the previous research were about 35%. Therefore, accuracy rate of the previous research is about 35% at a maximum. Fig. 3 shows the example of meaning estimation result.

In Fig. 3, relating to "AFC," all evaluation methods are successful in estimating right meaning. Next, relating to "NRA," only VSM fails in estimating right meaning. According to this result, it is considered that DoA and EMD which evaluate association in view of meaning of words function effectively compared to VSM which evaluates association by notation of words. Finally, relating to "HD," only EMD is successful in estimating right meaning. According to this result, it is considered that EMD which uses all attributes for calculation functions effectively compare to DoA which has the potential to eliminate partial attributes from calculation.

VII. CONCLUSION

This paper proposed the method to estimate the meaning of an alphabetical abbreviation. The proposal method regards meaning estimation of an alphabetical abbreviation as meaning estimation of unknown word. This method uses lexical conceptualization and evaluation approach for semasiological association. In addition, by using Wikipedia as a dictionary, this method complements lack of information

due to an alphabetical abbreviation. The method resolves the estimation of the original meaning. polysemy of alphabetical abbreviation and achieves

Alphabetical abbreviation (Meaning)	Number of meaning candidate	Input sentence	Evaluation method	Estimation result	Result
AFC (Asian Football Confederation)	13	... During 2008 to 2011 year, he was Federation Internationale de Football Association commissioner and AFC chairman, it's a reason that repeated ethical provision violations were found. ...	(1) DoA	Asian Football Confederation	○
			(2) EMD	Asian Football Confederation	○
			(3) VSM	Asian Football Confederation	○
NRA (National Rifle Association)	13	... Senator Joseph Manchin who is a member of lobby group NRA for gun owners and Second Amendment activist appeared on broadcast television on 17, and he made an appeal to regulate the lethal force firearm which is used in the current case. ...	(1) DoA	National Rifle Association	○
			(2) EMD	National Rifle Association	○
			(3) VSM	Nuclear Regulation Authority	×
HD (High Definition)	20	... On 12, America newspaper Wall Street Journal reported that America Apple is proceeding to design HD big-screen television with Asia electronics company such as Japan SHARP and Taiwan Hon Hai. ...	(1) DoA	Harley Davidson	×
			(2) EMD	High Definition	○
			(3) VSM	Harley Davidson	×

Fig. 3. The example of meaning estimation result.

The proposal method can make the meaning of an alphabetical abbreviation easier to understand, and it is consider that this method can contribute achievement of natural intellectual dialogue and performance improvement of many applications such as information search.

Possibility for future work is studying method for applying vector model which is based on neural network; as an example of such model, Word2Vec [14] is conducted active research in recent years. The very concept base can be potentially expanded (improved performance) by selecting and developing information resource for constructing vector model. Achievement of smooth natural language processing is expected by developing lexical conceptualization through such studying.

REFERENCES

[1] O. Naoaki and I. Mitsuru, "Abbreviation recognition in Japanese newspaper articles," in *Proc. the 21st Annual Conference of the Japanese Society for Artificial Intelligence*, 2G4-4, pp. 1-3, June 2007.

[2] M. Stevenson, Y. Guo, A. A. Amri, and R. Gaizauskas, "Disambiguation of biomedical abbreviation," in *Proc. the Workshop on BioNLP*, pp. 71-79, June 2009.

[3] R. Mihalcea, "Using Wikipedia for automatic word sense disambiguation," in *Proc. NAACL HLT*, pp. 196-203, April 2007.

[4] Y. Sun, L. Lin, D. Tang, N. Yang, Z. Ji, and X. Wang, "Modeling mention, context and entity with neural networks for entity disambiguation," in *Proc. the Twenty-Fourth IJCAI 2015*, pp. 1333-1339, July 2015.

[5] O. Noriyuki, T. Seiji, W. Hirokazu, and K. Tsukasa, "A construction of largescale concept-base for calculation of degree of association between concepts," *Journal of Natural Language Processing*, vol. 14, no. 5, pp. 41-64, Oct. 2007.

[6] Wikipedia. [Online]. Available: <https://jp.wikipedia.org>

[7] W. Hirokazu, O. Noriyuki, and K. Tsukasa, "The method of measuring the Degree of Association between concepts using attributes of the concepts and coincidence information," *Journal of Natural Language Processing*, vol. 13, no. 1, pp. 53-74, Jan. 2006.

[8] F. Yugo, W. Hirokazu, and K. Tsukasa, "Associative document retrieval using concept base and earth mover's distance," *Journal of Natural Language Processing*, vol. 16, no. 3, pp. 25-49, July 2009.

[9] T. Yasuki, W. Hirokazu, and K. Tsukasa, "The method of acquisition of the new concept and its attribute using the World Wide Web," in *Proc. the 18th Annual Conference of the Japanese Society for Artificial Intelligence*, 2D1-02, pp. 1-4, June 2004.

[10] Google. [Online]. Available : <http://www.google.co.jp>

[11] T. Takenobu, *Information Search and Language Processing*, Tokyo: University of Tokyo Press, 1999.

[12] A. J. Hoffman, "On simple linear programing problems," in *Proc. the Seventh Symposium in Pure Mathematics of the AMS*, vol. 7, pp. 317-323, 1963.

[13] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, Nov. 1975.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv Preprint*, pp. 1-12, Sep. 2013.



Kazuto Goto was born in Japan. He received the master's degree in engineering from Doshisha University of Kyotabae, Japan in 2008. He works as an engineer in Nippon Telegraph and Telephone Corporation of Yokosuka. His primary research interests are knowledge processing. Mr. Kazuto is membership of the Institute of Electronics, Information and Communication Engineers (IEICE).



Seiji Tsuchiya was born in Japan. He received the master's degree and doctor of philosophy degree in engineering from Doshisha University of Kyotabae, Japan in 2002 and 2007.

He worked as an engineer in SANYO Electric Company (2002-2007). He worked as an assistant professor in Tokushima University (2007-2009). Currently, he works as associate professor in Doshisha University of Kyotanabe. His primary research interests are knowledge processing and conceptual processing.

Assistant Prof. Seiji is membership of the Association for Natural Language Processing (ANLP), the Japanese Society for Artificial Intelligence, Information Processing Society of Japan (IPJSJ), Japanese Cognitive Science Society, and IEICE.



Hirokazu Watabe was born in Japan. He received the master's degree in engineering from Hokkaido University of Sapporo, Japan in 1985. He received the doctor of philosophy degree in engineering from Kyoto University of Kyoto, Japan in 1991.

He worked as an assistant professor in Kyoto University (1987-1994). He worked as a junior associate professor and an associate professor in Doshisha University (1994-1998 and 1998-2006). Currently, he works as a professor in Doshisha University of Kyotanabe. His primary research interests are evolutionary computation, computer vision, and conceptual processing.

Prof. Hirokazu is membership of ANLP, the Japanese Society for Artificial Intelligence, IPSJ, IEICE, the Institute of Systems, Control and Information Engineers, and the Japan Society for Precision Engineering.