A Semantic Inference Method of Unknown Words using Thesaurus based on an Association Mechanism

Seiji Tsuchiya*, Hirokazu Watabe**, Tsukasa Kawaoka**, Fuji Ren*
*The University of Tokushima, **Doshisha University
*, **Japan

1. Introduction

In recent years, a computer technology has advanced and a robot designed for home use is actively being developed. Home-entertainment robot like "AIBO"(Sony) and humanoid robot like "ASIMO"(Honda) which can walk on two feet are examples of such robot. It is not an exaggeration to say that hardware technology has advanced enough to fit for practical use at home.

Currently the field of artificial intelligence is expected to be developed as a software technology. For a robot to coexist with human, it should be equipped the ability to feel, think, talk or behave just like a person.

Therefore, we are engaged in research aiming to develop a robot which can smoothly make conversation with human beings. Such robot needs to have abilities to understand and interpret words. Currently, a technique based on a large-scale language dictionary or a corpus is predominantly used in the field of the language processing. Only one wrong response from a robot gives human beings very unfavourable impression during the conversation. This might become one of the fatal causes for human not to accept a robot.

Therefore, to achieve our purpose, a very large-scale language dictionary and a corpus are needed. As quite a lot of costs, resources and time are necessary to create such linguistic capital, automatic construction technique is also being researched. However, the knowledge in a category of common sense is inherent to human and difficult to construct automatically although it is indispensable knowledge for robot to realize conversation with human beings without sense of unease.

In this paper, we propose a technique which contributes to semiautomatic construction of a large-scale language dictionary and a corpus. Concretely, a system using the proposed technique indicates a position of an unknown word to be registered in an existing thesaurus dictionary. We show the effectiveness by comparing traditional techniques with the proposal technique. In addition, we evaluate how performance of the proposal technique approaches performance of human.

This system presents answer candidates so that a time and effort required for making a large-scale language dictionary and a corpus can be reduced.

2. Traditional Techniques

In this chapter, we explain techniques based on the vector space model and the statistical model. Then, these techniques and the proposed technique in this paper are compared, and the accuracy of an unknown word registration processing is evaluated.

2.1 Technique based on the Vector Space Model

In a technique based on vector space model, a similarity is calculated by using the cosine between a feature vector of each node in a thesaurus and a feature vector of an unknown word(Uramoto 1996). Then, the unknown word is registered in a node with a high similarity computation. In a simplest vector space model, a feature vector consists of a co-occurrence frequency of a noun and a verb. Each element of a feature vector at a node is calculated by adding co-occurrence frequencies of a verb and a noun at a node. Moreover, each element of a feature vector of an unknown word is a co-occurrence frequency of the unknown word and a verb.

$node_i$ shows a node of a thesaurus and $NODE, NODE = \left\{ node_1, node_2, \cdots, node_{|NODE|} \right\}$ is a node set with limited number of elements and $|\cdot|$ shows the number of the elements in a set. Moreover, $(w, z), w \in NODE, z \in VERB$ is a binomial class indicating one training data and means that node w and verb z are co-occurring. $(w, z)^N$ indicates series consisting of a training data of N pieces. $unknown$ meaning an unknown word and $(unknown, y^M)$ shows a binomial class of unknown word $unknown$ and series y^M of verb y co-occurring with unknown word $unknown$. When we define as above mentioned, in a technique based on vector space model, a node to register an unknown word is decided as follows:

$$d_{\cos}\left((w, z)^N, (unknown, y^M)\right) = \underset{node_i}{\arg\max} \cos(\, vec(node_i), vec(unknown)\,)$$

$$= \underset{node_i}{\arg\max} \frac{vec(node_i) \cdot vec(unknown)}{\| vec(node_i) \| \| vec(unknown) \|}$$

However,

$$vec(node_i) = \left(co\left((node_i, verb_1) \mid (w, z)^N \right), co\left((node_i, verb_2) \mid (w, z)^N \right), \cdots, co\left(\left(node_i, verb_{|VERB|}\right) \mid (w, z)^N \right) \right)$$

$$vec(unknown) = \left(co\left(verb_1 \mid y^M\right), co\left(verb_2 \mid y^M\right), \cdots, co\left(verb_{|VERB|} \mid y^M\right) \right)$$

Here, $d_{\cos}\left((w, z)^N, (unknown, y^M)\right)$ is a function to decide a node where an unknown word should be registered. $vec(A)$ shows a feature vector of *A* and $\cos(B \mid C)$ shows the number of *B* in *C*. Moreover, $\cos$ is a function which calculates a value of the cosine between vectors, $vec_A \cdot vec_B$ is inner product between vector vec_A and vec_B, and $\|vec\|$ is a norm of vector vec.

Besides a simple vector space model using co-occurrence frequency as mentioned above, vector space model using TF-IDF as a weight for each co-occurrence frequency has been proposed.

2.2 Technique based on the Statistical Model

Based on the statistical decision theory(Maeda, 2000), this technique minimizes an error rate that is a probability of registering an unknown word in a wrong node. This technique can be defined as follows:

$$d_{Bayes}\left(y^M\right) = \arg\max_{x \in NODE} \int_\Theta p\left(\theta \mid w^N z^N\right) p(x \mid \theta)\, d\theta \prod_{i=1}^{M} \int_\Theta p\left(\theta \mid w^N z^N, x, y^{i-1}\right) p\left(y_i \mid x, \theta\right) d\theta \quad (1)$$

Here, $p(\theta)$ shows a prior probability density function of parameter θ.

Moreover, the integration part can be transformed as below by assuming a beta distribution as a prior probability density function $p(\theta)$ of the parameter θ. $\beta(x)$ shows a parameter of a beta distribution corresponding to $p(x \mid \theta)$.

$$\int_\Theta p\left(\theta \mid w^N z^N\right) p(x \mid \theta)\, d\theta = \frac{co\left(x \mid w^N\right) + \beta(x)}{\sum_x \left(co\left(x \mid w^N\right) + \beta(x) \right)} \quad (2)$$

$$\int_\Theta p\left(\theta \mid w^N z^N, x, y^{i-1}\right) p\left(y_i \mid x, \theta\right) d\theta = \frac{co\left(xy_i \mid w^N z^N\right) + co\left(y_i \mid y^{i-1}\right) + \beta\left(y_i \mid x\right)}{\sum_{y_i} \left(co\left(xy_i \mid w^N z^N\right) + co\left(y_i \mid y^{i-1}\right) + \beta\left(y_i \mid x\right) \right)} \quad (3)$$

3. Proposed Technique

A proposed technique process an unknown word by evaluating the relevance between words using an Association Mechanism which has already been proposed. Concretely, semantic relation between word at a node of a thesaurus and an unknown word is evaluated with the Degree of Association then the unknown word is registered to a node with the closest relation.

An Association Mechanism consists of a Concept Base(Hirose et al., 2001)(Kojima et al., 2002) and the Degree of Association Algorithm(Watabe & Kawaoka, 2001). A Concept Base generates semantics from a certain word, and the Degree of Association Algorithm uses results of a expanded semantics to express the relation between one word and the other with a numeric value.

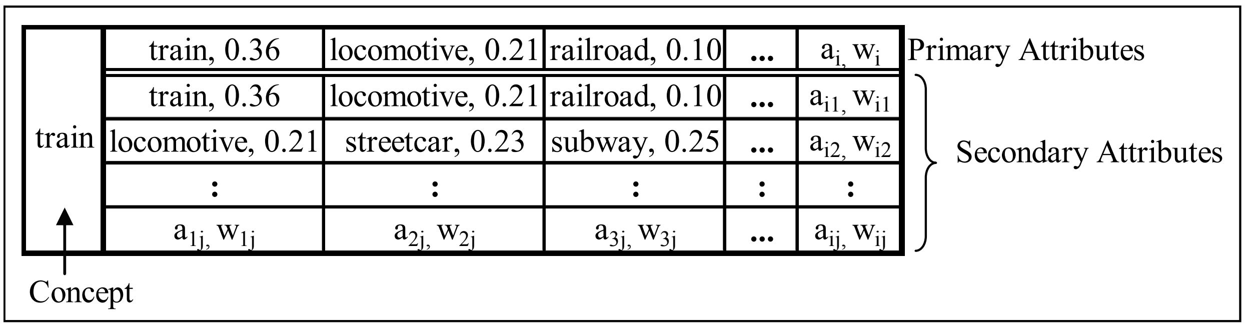
3.1 Concept Base

A Concept Base is a large-scale database constructed both manually and automatically from multiple electronic dictionaries. It has concept words, which are entry words taken from electronic dictionaries, and concept attributes, which are content words in the explanations of each entry word. In our research, a Concept Base containing approximately 90,000 concepts was used. The Concept Base went through auto refining process after the base had been manually constructed. In this processing, inappropriate attributes 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.

$$A = \{(a_1, w_1), (a_2, w_2), \ldots, (a_N, w_N)\} \tag{4}$$

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 1 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}	

Fig. 1. Example demonstrating the Concept "train" expanded as far as Secondary

3.2 Calculation 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:

$$\begin{aligned} A &= \{(a_1, u_1), (a_2, u_2), \ldots, (a_L, u_L)\} \\ B &= \{(b_1, v_1), (b_2, v_2), \ldots, (b_M, b_M)\} \end{aligned} \tag{5}$$

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):

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

The Degree of Association is obtained by calculating the Degree of Identity for all of the targeted Primary Attribute combinations and then by determining the correspondence between Primary Attributes. Specifically, priority is given to determine the correspondence between matching Primary Attributes. The correspondence between Primary Attributes that do not match is determined so as to maximize the total degree of matching. Using the degree of matching, it is possible to consider 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:

$$R(A,B)=\sum_{i=1}^{L} I(a_i,b_{xi})(u_i+v_{xi})\times\{\min(u_i,v_{xi})/\max(u_i,v_{xi})\}/2 \tag{7}$$

In other words, the Degree of Association 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. Unknown Word Registration Experiment

4.1 The Thesaurus used in This Experiment

A thesaurus is a dictionary where words are semantically classified and generally indicated with a tree structure. The thesaurus has two types: 1) a classification thesaurus with words only on leaf nodes and 2) a hierarchical thesaurus with words on root nodes and intermediate nodes besides leaf nodes.

In this paper, a hierarchical NTT thesaurus(NTT, 1997) was used for the experiment of the unknown word registration. Figure 2 shows a part of NTT thesaurus.

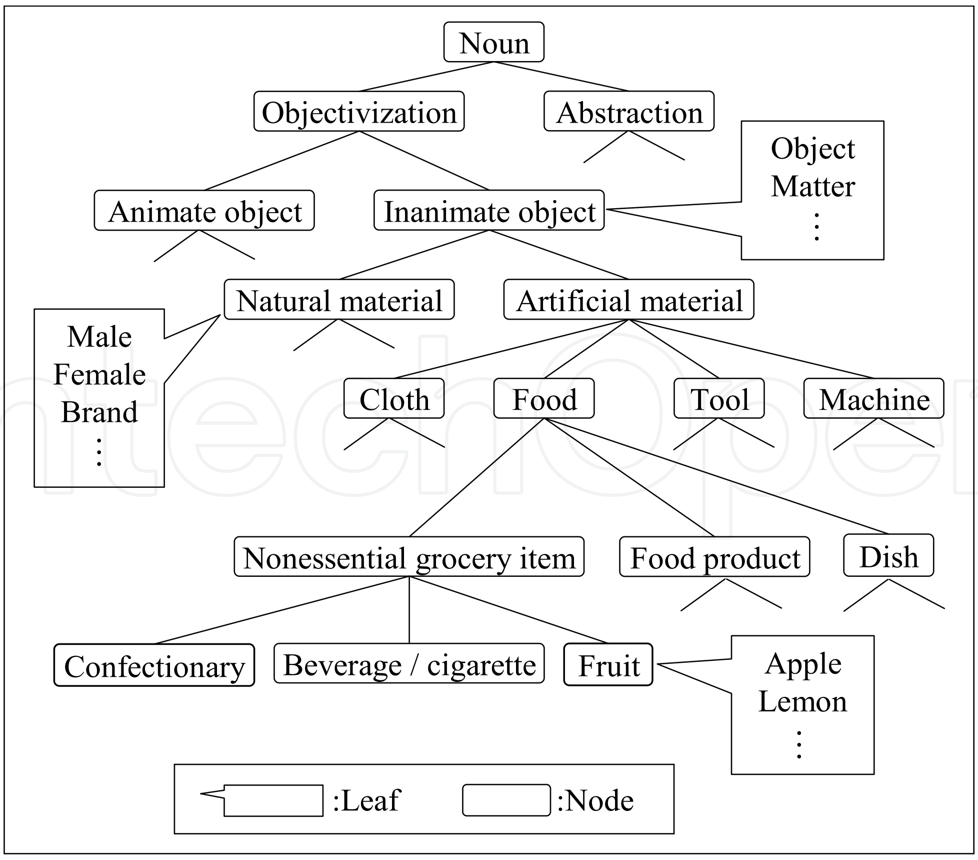


Fig. 2. An example of NTT thesaurus used in this paper

4.2 Method of Experiment

1000 words were extracted as unknown words from the words registered in NTT thesaurus explained in section 4.1. Two-stage sampling method was used as the extraction method. On the first stage, an equal probability sampling was carried out on the nodes with ten or more registered words. Then on the second stage, an equal probability sampling of non-restoration was carried out on noun words at the nodes. Examples of the words extracted as unknown words are shown in table 1.

Registered node	Unknown word
Remove	Diversion
Inversion	Opposition
Rejection	Declination
Union	Bridal
Price	Advance
Woman	Girl
Dropping	Drip
Superior	Seniority
Detainment	Captivity
City	Suburb

Table 1. Examples of unknown words

4.3 Evaluation

We evaluated each technique according to the correct answer rate of the top 10 candidate nodes to register an unknown word. Here, the node where the unknown word is registered in NTT thesaurus is considered as correct answer. When the unknown word is registered in two nodes or more, we judge the answer correct if the outputted node matches one of the registered nodes.

A result of an unknown word registration experiment is shown in figure 3. The axis of abscissas in figure 3 is the number of the considered accumulative candidates and the spindle is the correct answer rate (rate of accuracy). “Cos” is a vector space model using only co-occurrence frequency, “TF-IDF” is the vector space method using TF-IDF for a weight of co-occurrence frequency, “Bayes” is a technique using statistical model (Bayes theory) introduced in section 2.2 and “DA” is proposed technique in this paper.

In addition, the accuracy of this unknown word registration by human is approximately 89.4%.

4.4 Discussion

Figure 3 shows that the proposed technique is generally better than the traditional techniques. When the first answer was outputted, the accuracy improved only approximately 4%. But the accuracy improved approximately 20% if the top five answers or more were output. This result suggests that the proposed technique which semantically extends a word using the Concept Base and evaluates a semantic relation between words using the Degree of Association should be able to understand a vocabulary efficiently than the traditional techniques based on the probability and statistics.

As described in section 4.3, the accuracy when human solves the test used in this paper is approximately 89.4%. So, when human's accuracy is considered 100%, the accuracy of the first answer outputted by proposed technique is approximately 45.9% and the accuracy of the top 10 answers outputted by proposed technique is approximately 82.4%. Thus, when the top four answers or more are outputted, performance of proposed technique approaches performance of human by approximately 70% or more.

The main purpose of this paper is to construct a large-scale language dictionary and a corpus not automatically but semi-automatically. Therefore we think that it is more important to have the correct answer efficiently included in two or more answer candidates than in the first answer. So, it is considered that the proposed technique in this paper is a very effective technique.

However, we do not think that the accuracy of the first answer is enough. The proposed technique calculates the Degree of Association between an unknown word and a node of a thesaurus using only words at the node. In the future, we would like to improve the accuracy by a new registration method which uses registered nodes and leaf nodes extending from the node for the calculation of Degree of Association between an unknown word and a node in a thesaurus. Moreover, in this paper an unknown word was pseudo made from the word registered in an existing thesaurus. However, in the future, we would like to conduct a similar experiment which uses a true unknown word not registered in an existing thesaurus and the Concept Base.

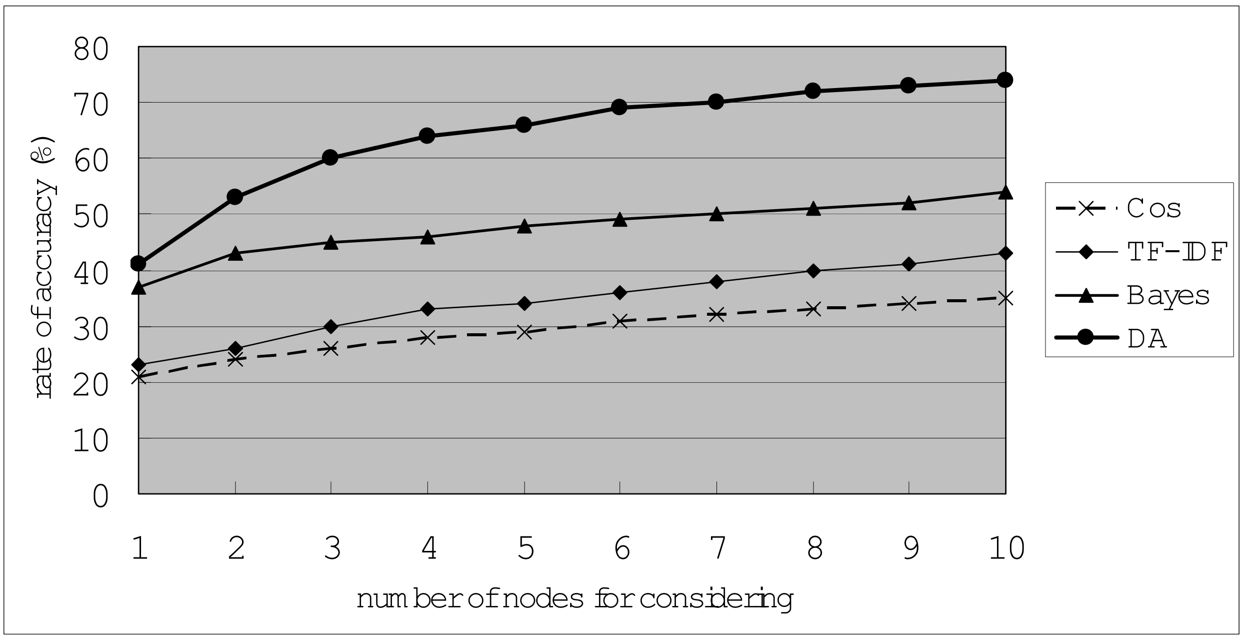


Fig. 3. A result of an unknown word registration experiment

5. Conclusion

In this paper, to reduce costs, resources and time, we proposed a technique which semi-automatically constructs a large-scale dictionary and corpus by using an Association Mechanism based on the Concept Base and the Degree of Association.

The proposed technique was able to improve the accuracy approximately 20% as a result compared with the traditional techniques.

In addition, when the top four answers or more were outputted, performance of proposed technique approached performance of human by approximately 70% or more.

References

Sony. http://www.sony.jp/products/Consumer/aibo/

Honda. http://www.honda.co.jp/ASIMO/

Uramoto, N. (1996). Corpus-based Thesaurus – Positioning Word in Existing Thesaurus Using Statistical Information from a Corpus, Journal of Information Processing Society of Japan, vol.37, No.12, pp.2182-2189.

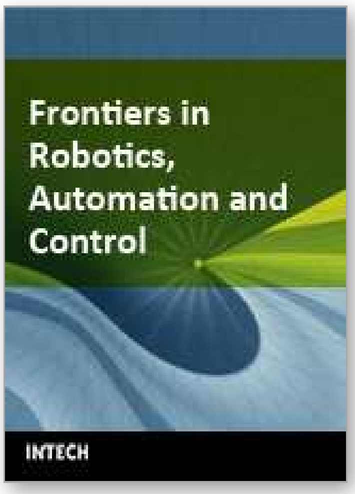
Maeda, Y. (2000). A Note on Positioning Unknown Words in an Existing Thesaurus Based upon Statistical Decision Theory, The IEICE transactions on information and systems (Japanese edition), Vol.J83-A, No.6, pp.72-710.

Hirose, T.; Watabe, H. & Kawaoka, T. (2002). 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.

Kojima, K.; Watabe, H. & Kawaoka, T. (2002). 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*, *9*(5), pp.93-110.

Watabe, H. & Kawaoka, T. (2001). Measuring Degree of Association between Concepts for Commonsense Judgements, *Journal of Natural Language Processing*, *8*(2), pp.39-54.

NTT Communication Science Laboratory. (1997). *NIHONGOGOITAIKEI*, Iwanami Shoten, ISBN4-00-009884-5 C3581.



Frontiers in Robotics, Automation and Control
Edited by Alexander Zemliak

ISBN 978-953-7619-17-6
Hard cover, 450 pages
Publisher InTech
Published online 01, October, 2008
Published in print edition October, 2008

This book includes 23 chapters introducing basic research, advanced developments and applications. The book covers topics such us modeling and practical realization of robotic control for different applications, researching of the problems of stability and robustness, automation in algorithm and program developments with application in speech signal processing and linguistic research, system's applied control, computations, and control theory application in mechanics and electronics.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821