Viewpoint-Based Measurement of Semantic Similarity between Words

Kaname KASAHARA, Kazumitsu MATSUZAWA, Tsutomu ISHIKAWA and Tsukasa KAWAOKA

NTT Communication Science Laboratories,
1-2356 Take, Yokosuka-shi, Kanagawa, 238-03, Japan
kaname@nttkb.ntt.jp

Abstract

A method of measuring semantic similarity between words using a knowledge-base constructed automatically from machine-readable dictionaries is proposed. The method takes into consideration the fact that similarity changes depending on situation or context, which we call 'viewpoint'. A feature of the method is that certain parts of the overall concept of words, compared with each other, are emphasized by using the viewpoint when calculating the degree of similarity. Evaluation shows the proposed method, although based on a simply structured knowledge-base, is superior to other currently available methods.

1. Introduction

Measuring semantic similarity between words is important for natural language processing in text-search, analogical reasoning, cases-based reasoning, flexible human interfaces to databases, etc. We research methods for measuring the similarity between large numbers of daily-use words with an aim toward general applications. In measuring such similarity, it must be taken into consideration that similarity changes depending on situation or context, which we call 'viewpoint'. For example, 'horse' is more similar to 'pig' than 'car' from the viewpoint of 'animal'. On the other hand, 'horse' is more similar to 'car' from the viewpoint of 'vehicle'. Babaguchi et. al. proposed measuring similarity between cases in a database based on such a viewpoint[1]. However, they

assumed the attributes of the cases are already known, since their purpose was application to an existing database. Therefore it is difficult to apply their method to measuring similarity between the daily-use words themselves used in the aforementioned applications.

In this paper, we propose a method that takes viewpoint into consideration in measuring the similarity between words. The method uses a knowledge-base which is constructed automatically from machine-readable dictionaries. Certain parts of the overall concept of words compared to each other, which are common to the knowledge of the viewpoint, are emphasized when calculating degree of similarity.

We built an experimental knowledge-base and evaluated this method with respect to its preciseness in similar word retrieval. The results obtained show our method is superior in this regard to the conventional one based on a tree-structured thesaurus.

2. Method for measuring similarity between words

2.1. The knowledge-base of words

Knowledge of a word is typically represented by a series of lists, with each list consisting of the attribute of the word and the value of the attribute. For example, the word 'apple' may be represented as

$$'apple' = \{('shape','sphere'),\ ('color','red'), ('taste','sour'), \cdots\}.$$

During the past several years, the CYC[2] and EDR[3] projects have been attempting to acquire an enormous amount of such semantic knowledge. However, it is difficult to obtain all the attributes and values for a large number of daily-use words by hand from the quantitative point of view[4]. On the other hand, machine-readable dictionaries have been seen as a likely source of semantic knowledge, and considerable research on way of extracting the knowledge from them is still going [5, 6]. However, no adequate method for acquiring the attributes and values completely and fully has been found yet. Moreover, the viewpoint is not considered in any of the research projects referenced above. Therefore, we constructed a simply structured knowledge-base that can be made automatically from machine-readable dictionaries only, and tried to attain judgement of viewpoint-based similarity.

In our method, each word $Word_i$ in the knowledge-base consists of a series of lists, with each list consisting a keyword p_{ij} of $Word_i$ and a weight q_{ij} of p_{ij} where

$$Word_i = \{(p_{i1}, q_{i1}), (p_{i2}, q_{i2}), \cdots, (p_{in}, q_{in})\}. \quad (1)$$

We choose the morpheme contained in the definition or explanation of the word in the dictionary as the keyword of the word, and choose the number of times that the morpheme appears as the weight of the keyword.

2.2. Similarity measurement procedure

The degree of similarity S is calculated using only lists of the keywords and weights of the words included in the knowledge-base (Figure 1).

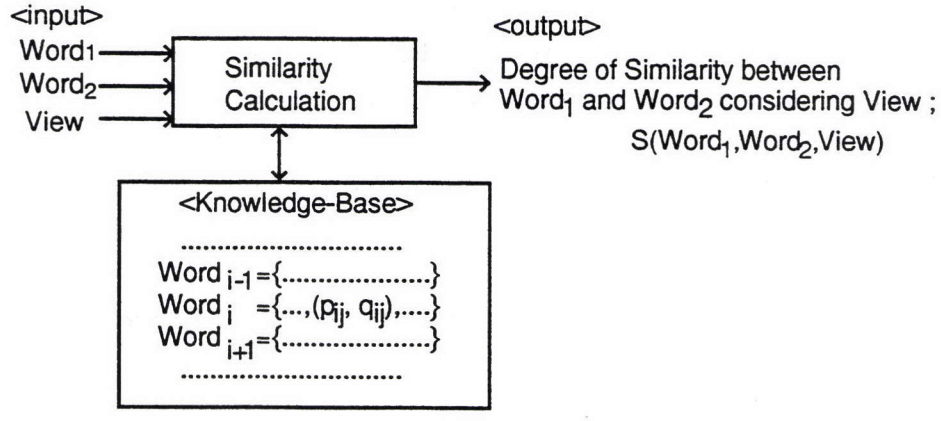

Figure 1: A similarity measurement scheme

Here, $Word_1$ and $Word_2$ are words between which the degree of similarity is calculated, and $View$ is the word which represents the viewpoint (we call it the 'viewpoint word'). $S(Word_1, Word_2, View)$ is calculated as per the following steps:

- **STEP 1: Standardization of the keyword**
  Keywords which are semantically similar to each other should be identified, because the measurement of the similarity is basically calculated by comparing the keywords of words. We call this identification 'standardization'. Here, each keyword is standardized by placing it into the category of a thesaurus.

- **STEP 2: Normalization of the weight**
  The number of the keyword and the value of its weight depend on the length of the definition or explanation about the word in the dictionary. That is, words in the knowledge-base have different numbers of positive weights values. Therefore, all weights should be normalized so that their degree of similarity is independent of the differences between them.

- **STEP 3: Modulation based on viewpoint**
  When the degree of similarity based on the viewpoint is calculated, we assume that categories in the viewpoint word which have a positive weight are important. Therefore, certain part of weights of a word are emphasized as a function of the

weight of the same category contained in the viewpoint word when the degree of similarity is calculated. We call this emphasis 'modulation'.

- **STEP 4: Calculation of the degree of similarity (S)**
  We consider S as the simple degree of similarity between two modulated words. Accordingly, S is defined based on the idea of the angle between two vectors representing the modulated words.

2.2.1. Keyword Standardization

The degree of similarity is basically calculated by comparing keywords. Therefore, keywords which are nearly equal in meaning should be regarded as being the same keywords. For example, the keyword 'shape' should be equal to the keyword 'form'. We use the the thesaurus T consisting of k categories for this standardization. c_m is the category in which all the words p_{ml} in the set are similar to each other.

$$T = \{c_1, c_2, \cdots, c_m, \cdots, c_k\}$$
$$c_m = \{p_{m1}, p_{m2}, \cdots p_{ml}, \cdots\}. \tag{2}$$

We regard the categories in T as being independent of each other from the semantic point of view. In this case, the categories of T span vector space V of k dimensions.

Each keyword of $Word_i$ is standardized by placing it into the category in which it is included. Using these standardized keywords, vectorized word $\mathbf{Word_i}$ is generated as follows:

$$\mathbf{Word_i} = (Q_{i1}, Q_{i2}, \cdots, Q_{ij}, \cdots, Q_{ik})$$
$$Q_{ij} = \sum_{l=1 \ p_l \in c_j}^{n} q_{il}. \tag{3}$$

Here, the second equation shows that the weights of the keywords are summed up, if there are some keywords included in the same category.

2.2.2. Normalization of the weight

Because the number of the keyword and the value of its weight depend on the length of the definition or explanation of the word in the dictionaries, the value and number of positive weights vary with the word in the knowledge-base. When calculation of the similarity is based on the distance between $\mathbf{Word_1}$ and $\mathbf{Word_2}$ in V, the value and number of positive weights dominantly lead to decide the degree of the similarity. For example, words which have only a small number of small positive weights are always nearer and more similar to each other than to a word which has a large number of large positive weights. Therefore, the weights of the word should be normalized so that the

length of $\mathbf{Word_i}$ is constant. We normalize the weight of $\mathbf{Word_i}$ by dividing it by the the vector norm $\|\mathbf{Word_i}\|$ of the $\mathbf{Word_i}$:

$$\mathbf{Word_i} = (\hat{Q}_{i1}, \hat{Q}_{i2}, \cdots, \hat{Q}_{ij}, \cdots, \hat{Q}_{ik})$$
$$\hat{Q}_{ij} = \frac{Q_{ij}}{\|\mathbf{Word_i}\|} = \frac{Q_{ij}}{\sqrt{\sum_{m=1}^{k} Q_{im}^2}}. \tag{4}$$

2.2.3. Modulation based on viewpoint

To take the viewpoint into consideration in calculating the degree of similarity, words compared with each other should be emphasized using the viewpoint word. The viewpoint word (**View**) is also standardized and normalized as:

$$\mathbf{View} = (\hat{Q}_{v1}, \hat{Q}_{v2}, \cdots, \hat{Q}_{vj}, \cdots, \hat{Q}_{vk}) \qquad (\|\mathbf{View}\| = \mathbf{1}). \tag{5}$$

We assume that categories which have positive weight in **View** should be important when the degree of the similarity is calculated. Therefore, we define 'modulation' as the procedure in which the weights of such categories in the $\mathbf{Word_i}$ are emphasized when calculating the degree of similarity. A modulated word $\mathbf{Word_i^v}$ is generated as per the following formula:

$$\mathbf{Word_i^v} = (Q_{i1}^v, Q_{i2}^v, \cdots, Q_{ij}^v, \cdots, Q_{ik}^v)$$
$$Q_{ij}^v = \hat{Q}_{ij} \cdot M(\hat{Q}_{vj}). \tag{6}$$

Because all categories in T are regarded as being independent of each other, the weights of $\mathbf{Word_i^v}$ Q_{ij}^v can be calculated only from weights of the same category in $\mathbf{Word_i}$ and **View**. M is a function of modulation, which decides the degree of emphasis for each category. We assume that small weights in the viewpoint are the noise weights, which have no relation with the viewpoint. Therefore, we adopt the modulation function shown below:

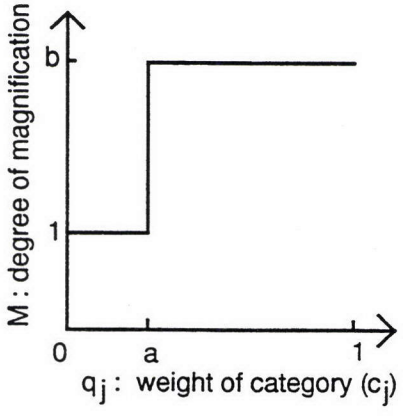

Figure 2: Modulation function example

In Fig 2, a is the threshold value which discerns noise weight. Next the modulated word $\mathbf{Word}_i^{\mathbf{v}}$ is normalized again by Equation 4, because the length of the vector norm is changed by modulation.

2.2.4. Calculation of the degree of similarity (S)

We calculate $S(Word_1, Word_2, View)$ as the similarity between two words modulated.

$$S(Word_1, Word_2, View) = R(\mathbf{Word}_1^{\mathbf{v}}, \mathbf{Word}_2^{\mathbf{v}}). \tag{7}$$

R is the function which indicates the nearness between two modulated words in the vector space V. We consider R to require the following conditions:

- $0 \leq R(\mathbf{Word_a}, \mathbf{Word_b}) \leq 1$
- $R(\mathbf{Word_a}, \mathbf{Word_b}) \equiv R(\mathbf{Word_b}, \mathbf{Word_a})$
- $R(\mathbf{Word_a}, \mathbf{Word_a}) \equiv 1$
- $\mathbf{Word_b}$ is more similar to $\mathbf{Word_c}$ than $\mathbf{Word_a}$,
  if $R(\mathbf{Word_a}, \mathbf{Word_b}) \leq R(\mathbf{Word_c}, \mathbf{Word_b})$.

We select the cosine of the angle θ between $\mathbf{Word}_1^{\mathbf{v}}$ and $\mathbf{Word}_2^{\mathbf{v}}$ as R, because the length of the modulated vector is constant. This satisfies the above mentioned conditions.

$$\begin{aligned} S &= cos\theta = \mathbf{Word}_1^{\mathbf{v}} \cdot \mathbf{Word}_2^{\mathbf{v}} \\ &= \sum_{j=1}^{k} Q_{1j}^{v} \, Q_{2j}^{v}. \end{aligned} \tag{8}$$

3. Evaluation

3.1. Experimental knowledge-base

We made an experimental knowledge-base of 40,000 Japanese daily-use words using four Japanese dictionaries [8, 9, 10, 11] to evaluate the proposed method. To achieve the standardization described in section 2.2.1, we use a thesaurus [7] containing 370,000 words grouped into 3,000 categories. Each word in the constructed knowledge-base has on average fifty lists, each consisting of a keyword and its weight.

3.2. Examples of judgement of similarity

In the following example, we calculate the degree of viewpoint-based similarity of the word, 'water' and 'oil':

Table 1: A degree of similarity between 'water' and 'oil'

viewpoint-word *View*	Degree of similarity S('water', 'oil', *View*)
(no word)	0.30
'molecular'	0.24
'liquid'	0.48

This table shows that the degree of similarity between the two words based on an inadequate viewpoint word ('molecular') is even smaller than the degree calculated when no viewpoint word is given at all. On the other hand, if an adequate viewpoint word ('liquid') is selected, the similarity is considerably larger than that calculated without a viewpoint word. Given the viewpoint word 'molecular', a person might not answer that 'oil' is similar to 'water'. Given the viewpoint word 'liquid', however, the person very likely will answer that 'oil' is similar to 'water'. In this case, then, the judgement of Table 1 may appears to be "human-like".

Another example follows in Table 2.

Table 2: An example of selecting a word similar to 'silver'

viewpoint-word *View*	Degree of similarity	
	S('silver', 'gold', *View*)	S('silver', 'lead', *View*)
metal	0.82	0.81
cash	0.89	0.24

The degree of similarity between 'gold' and 'silver' is as large as that between 'lead' and 'silver' when 'metal' is selected as a viewpoint word, which means both 'lead' and 'gold' are similar to 'silver'. On the other hand, 'silver' is much more similar to 'gold' than 'lead' is when 'cash' is selected as the viewpoint word. This result, too, appears to be very "human-like".

When the degrees of similarity between an indicated word and each of the 40,000 words in the knowledge-base are calculated, retrieval of words which are similar from the given viewpoint can be done by listing words in the order of their degree of similarity. Table 3 is a list of the highest-ranking words which are similar to the Japanese word "banshū"(=late fall) from the viewpoint of "owari"(=the end of something; an ending), as obtained by using our method as a retrieval system.

Table 3: An example of related word retrieval

Rank	Related words	Related words (translated into English)	Degree of Similarity to "ban-shū" from the viewpoint of "owari")
1	banshū	late fall	1.000
2	oshimai	conclusion	0.784
3	banshun	late spring	0.772
4	bantō	late winter	0.771
5	banka	late summer	0.765
6	yūkoku	evening	0.744
8	boshun	late spring	0.739
9	higure	sunset	0.730
10	yūmagure	evening	0.729
11	kure	end of the year	0.727
12	shijū	always	0.723
13	hitomoshigoro	evening	0.721
14	yagate	soon	0.715
15	tettōtetsubi	thoroughly	0.714
15	shoshū	early fall	0.714
17	tasogare	dusk	0.707
18	nenmatsu	year-end	0.706
19	yūbe	evening	0.702
20	tomeru	stop	0.701

Most of the words retrieved in Table 3 are similar to the given word ('banshū'). However, some words are not similar (for example, 'shijū', 'yagate' and 'tettōtetsubi'). In the next section, therefore, we evaluate our method through several lists of words such as that given above.

3.3. Evaluation of the proposed method

We evaluated measurement of similarity with the parameter F $(0 < F \leq 1)$, which is calculated using the precision P_i used for evaluation in information retrieval. These values are given as follows:

$$F = \frac{1}{n}\sum_{i=1}^{n} P_i$$
$$P_i = \frac{A \cap B_i}{B_i}. \tag{9}$$

When a word and the viewpoint-word are given, A is the set of n similar words selected by humans , and B_i is the minimum set of similar words which are selected by the retrieval system and which contains i words of A as Figure 3. The system is regarded as good when F is near to 1.

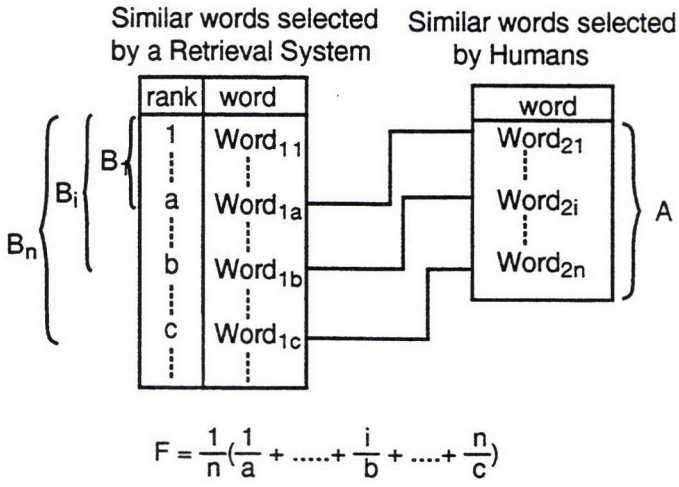

Figure 3: A scheme of calculating F

Next, let us consider a way of comparing the proposed method with another method. As a subject for comparison, we selected a conventional method which calculates the degree of similarity based on the distance between words in the thesaurus, because a thesaurus is generally thought to be a tool for judging similarity. We use the same tree-structured thesaurus which is used for standardization, and define the distance L as the number of categories which exist on the path between two categories, each of which includes words compared with each other. We define the degree of similarity S' in a conventional manner using L as follows:

$$S' = 1 - \frac{L}{L_{max}}. \tag{10}$$

Here L_{max} is the maximum distance between every two categories in the thesaurus, and S' satisfies the degree of similarity function conditions mentioned in Section 2.2.4.

We show an evaluation example in Figure 4. In this figure, F for the proposed method and the conventional method are compared for fifty-eight samples.

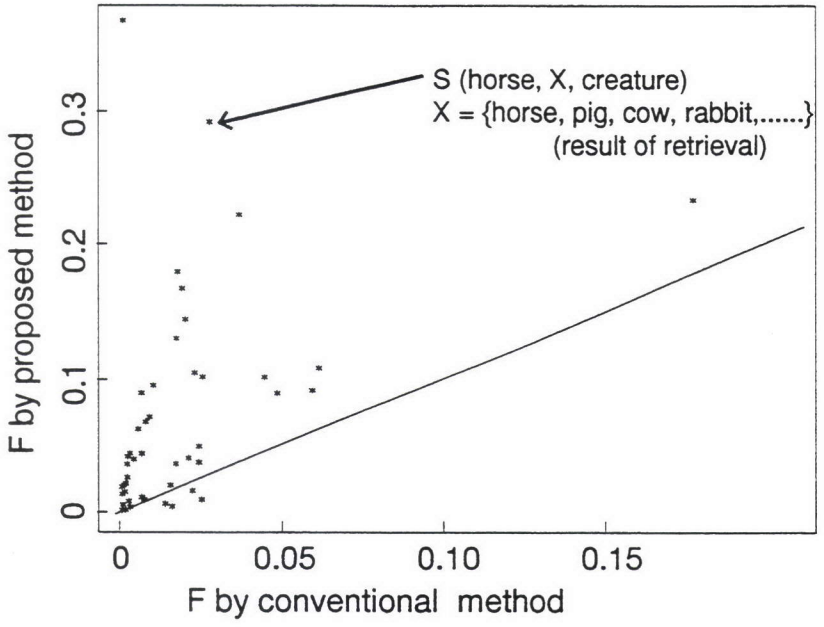

Figure 4: Result of evaluation

Each point in Figure 4 means that evaluations of retrievals based on our method and the conventional method are done with a given word and a viewpoint word. The value of the horizontal axis of the point is F calculated based on our method , and the value of the vertical axis on it is F calculated based on the conventional method . From this figure, the value of F based on our method is larger than that of the conventional method about 80% of samples. Morever, the ratio of F based on the proposed method to F based on the conventional one at each point is about fifteen to one on average. These facts mean that our proposed method is superior to the conventional one which adopted the thesaurus directly.

4. Conclusion

We have proposed a method for viewpoint-based measurement of the semantic similarity between words based on a knowledge-base constructed automatically from machine-readable dictionaries. We constructed the experimental knowledge-base of 40,000 Japanese words and evaluated our method in terms of the efficiency of retrieval of similar words. Test results confirmed that the proposed method, although based on the simply structured knowledge-base, is superior to the conventional one.

References

[1] Noboru Babaguchi, Yuji Sawada and Takenao Ohkawa, "Association mechanism based on viewpoints", *Transactions of Information Processing Society of Japan*, 35(5), pp.714–724, 1994.

[2] R.V.Guha and D.B.Lenat, "Cyc: A Midterm Report", AI Magazine, 11(3), pp.32–59, 1990.

[3] Japan Electronic Dictionary Research Institute, Ltd., "EDR Electronic Dictionary Technical Guide", TR-042, 1993.

[4] Nancy Ide and Jean Veronis, "Extracting Knowledge Base from Machine-readable Dictionaries: Have we wasted our time?", *Proceedings of the First International Conference on Building and Sharing of Very Large-Scale Knowledge Bases*, pp.257-266, 1993.

[5] Nancy Ide and Jean Veronis, "Combining Dictionary-Based and Example-Based Methods for Natural Language Analysis", *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, pp.69–78, 1993.

[6] Jun-ichi Nakamura and Makoto Nagao, "Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation", *Proceedings of the 13th International Conference on Computational Lingusitics*, pp.459–464, 1988.

[7] Satoshi Ikehara, Masahiro Ikehara, and Akio Yokoo, "Classification of language knowledge for meaning analysis in machine translations", *Transactions of Information Processing Society of Japan*, 34(8):1692–1704, 1993.

[8] Shinmura Izuru, editor, "Koujien", Iwanami Shoten, 1991.

[9] Akira Matsumuura, editor, "Daijirin", Sanseido Co.,Ltd., 1992.

[10] Masando Hamanishi and Susumu Ono, "Ruigo Kokugo Jiten", Kadokawa Shoten, 1990.

[11] et.al Takashi Ichikawa, editor, "Sanseidou genndai Kokugo jiten", Sanseido Co.,Ltd., 1992.