Comparison of two weights of relation between words: recollection weight and relevancy weight

Kaname Kasahara
NTT Communication Science Laboratories,
Nippon Telegraph and Telephone Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun,
Kyoto 619-0237, Japan
kaname@cslab.kecl.ntt.co.jp

Nozomu Inago
NTT Communication Science Laboratories,
Nippon Telegraph and Telephone Corporation
inago@cslab.kecl.ntt.co.jp

Tomoko Kanasugi
NTT Advanced Technology Corporation
kanasugi@cslab.kecl.ntt.co.jp

Chiharu Nagamori
NTT Advanced Technology Corporation
chiharu@nlp.ntt-at.co.jp

Tsuneaki Kato
Language and Information Sciences, The University of Tokyo
3-8-1 Komaba, Meguro-ku Tokyo, 153-8902, Japan
kato@boz.c.u-tokyo.ac.jp

ABSTRACT

In order to analyze how to construct a standard database of human judgments about word relations, two sorts of weights of relation between words, recollection and relevancy, are compared to each other and discussed based on the results of two experiments employing human subjects. We selected synonymy, antonymy, and associativity as the relations to be tested and found that there were basically middle degrees of correlation between the two sorts of weight in synonymy and antonymy. However, there was a low degree of correlation in associativity. This suggested that there are more words which can be judged to be relevant associative to a stimulus word than ones which are simply recollected.

Keywords: synonymy, antonymy, associativity, semantic similarity, thesaurus

1. Introduction

Simulations of semantic judgments about words, such as retrieving associative words and clustering words based on similarity (for example, [13, 10]) are useful for realizing intelligent text processing technologies like information retrieval[5, 14], machine translation[8], and text segmentation[3]. When studying a simulation method, how close the result of the method is to human judgment needs to be evaluated.

Usually, a large-scale database of human judgments, which contains word pairs and names of relations that connect words in the pairs, is used for that purpose. In English, WordNet[6] is one well-known database and there are Nihongo-GoiTaikei[7] and EDR[9] in Japanese. These databases were constructed based on the knowledge of a few linguists or psychologists. Therefore, they contain words and relations which even native speakers of a language might rarely use in their lifetime. Moreover, information about the weight of relation between words, which represents how strongly the words are connected in a given kind of relation such as synonymy, is lacking in the databases. This weight information is important when a database of human judgments about words is used to get the most related word to a given word or to get a ranked word list based on the weights of relations between a stimulus word and target words when a method of simulating related word retrieval is evaluated.

In psychology studies, several norms of word association [11, 12, 15], were built through human subjects experiments in which the free association method was used. They contain weights of associated words, which are estimated based on how many subjects recollected the same words. This weight information can be applied to evaluate the simulation method. We call this weight “recollection weight.” Meanwhile, another sort of weight for a relation is required to compare human judgment of whether two given words are related or not with a simulation method. The weight is estimated from how many subjects judge that the target word has a relevant relationship with a stimulus word. We call this "relevancy weight."

If there is always a high degree of correlation between the recollection weight and the relevancy weight in any kind of relation, an experiment to acquire only one of the two weights is sufficient to build a database of human judgments about relations between words. However, an analysis of how strong there is correlation has not yet been made.

In this paper, we report the result of analyzing correlation between two sorts of weight in the relations synonymy, antonymy, and associativity acquired from two human subject experiments for stimuli of 200 Japanese words in daily use. We also show the effectiveness of the standard database made through these experiments by comparing it with a large-scale thesaurus and by applying it to evaluate the method which simulates judgment of synonymy between words by employing a machine-readable dictionary[10].

2. Experiments

In this section, details of two human subject experiments for getting recollection and relevancy weights between stimulus and target words which are highly related to each other are described.

Stimulus words

Stimulus words should be selected from commonly known words because all the subjects participating in the experiments should be familiar with the stimuli. Usually, frequently appearing words in a text corpus such as newspapers and novels are regarded as familiar words. However, it is well known that word frequency does not always reflect how common the word is. Therefore, we used the database of word familiarity[2] to select commonly known Japanese words as stimuli. The familiarity database was developed for about 80,000 Japanese words whose familiarity scores in the database were measured by 32 Japanese adults using a seven-point scale.

Two hundred stimulus words were randomly selected from 28,764 words whose familiarity scores are more than five and which were estimated to be commonly known by more than 90% of Japanese adults[1].

Table 1. Examples of stimuli

"kuma"(bear), "kenchiku"(construction), "shouhi"(consumption), "hikohki"(airplane), "sakusen"(operation), "odoru"(dance), "youshoku"(western food), "uma"(horse), "hamusutah"(hamster), "taigah"(tiger), "hikkoshi"(removal), "tsuku"(arrive), "bakushou"(explosion of laughter), "matsuri"(festival)

Kinds of relation

It has been thought that there are several kinds of relation between two words. For example, in WordNet[6], seventeen relations such as HYPERNYM and SYNSET are listed. As the first step of our study, synonymy, antonymy, and associativity, which seem to be most directly connected to studies of simulating judgment of semantic similarity between words, were selected as the relations to be tested to get two sorts of weights for them.

Experiment 1

The first experiment was done to get recollection weights for synonymy, antonymy, and associativity.

One hundred adult subjects were asked to write down recollected words for each of the 200 stimulus words on questionnaire sheets for one of the three given relations. They were required to write down as many words as they could in ten seconds for one stimulus word. For each stimulus word and one of the target words recollected by any subject for one of the relations, the frequency of how many subjects wrote the same target word was counted as recollection weight. As a result, on average, about 40 target words and their recollection weights of synonymy, 33 target words and their recollection weights of antonymy, and 128 words and their recollection weights of associativity were acquired using one of the stimuli.

Experiment 2

For each stimulus word and its target word recollected in Experiment 1, seventy six subjects were asked to judge relevancy, namely, whether there was a given relation between the stimulus word and the target word. The frequency of how many subjects judged there was relevancy between a given stimulus word and target word was regarded as the relevancy weight between the words.

As shown in Table 2, in any relation, number of the target words whose recollection weights were only one are more than half of all the target words. In order to shorten time of the second experiment, 600 target words of the 2,909 whose recollection weights were one in synonymy, 596 target words of the 2,229 in antonymy, and 8,215 target words of the 17,418 in associativity were selected to score the relevancy weight.

Table 2. Number of acquired recollected words in Experiment 1

relation	weight of recollection		
	all	one	more than two
synonymy	7,988	5,079	2,909
antonymy	6,514	4,285	2,229
associativity	25,633	17,418	8,215

3. Results

Analyses of correlation between two sorts of weight for relations

Table 3 shows the correlation coefficients between recollection weight and relevancy weight in synonymy, antonymy, and associativity. These figures suggest that the two sorts of weight have a middle degree of correlation in synonymy and antonymy. In associativity, however, the degree of correlation is low.

Table 3. Correlation between recollection weight and relevancy weight

relation	correlation coefficient
synonymy	0.465
antonymy	0.418
associativity	0.294

Figures 1 to 3 show correlation between the recollection weights and the averaged relevance weights of word pairs that have the same recollection weight in the three relations. These figures reveal that only in the relation of associativity, are relevancy weights fairly large even when recollection weight is small. When several words were recollected by only one of the 100 subjects, these words were judged, on average, to be relevant and to have associative relations by more than half of the other 76 subjects on average.

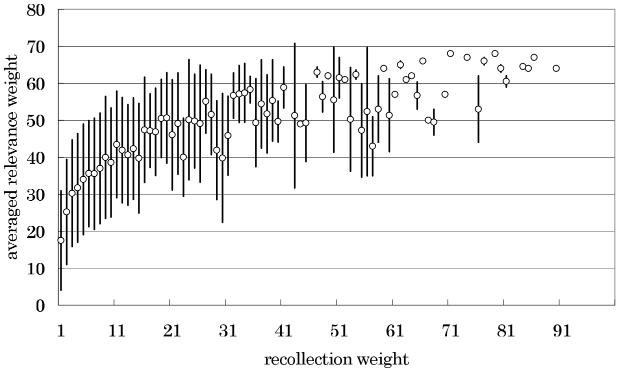

Figure 1. Correlation in synonymy

This result suggests that there are many possibly relevant words other than recollected words in the relation of free associativity only. To confirm this inference, we made pairs of a stimulus word of the 200 and a non-recollected word randomly selected from all the answered target words in experiment 1, 600 in synonymy, 600 in antonymy, and 1,000 in associativity. In Experiment 2, these pairs were tested simultaneously. The distributions of relevancy weights for the non-recollected pairs in one of the three relations are shown in Figure 4.

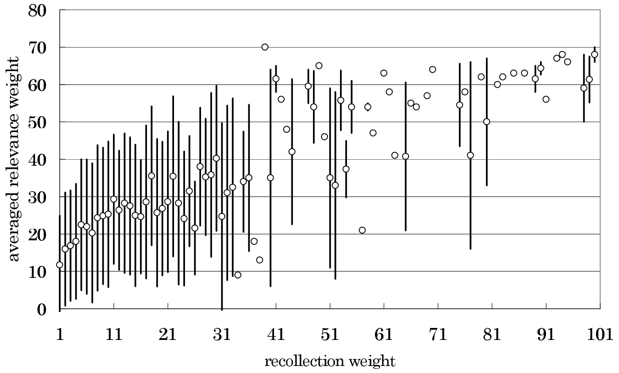

Figure 2. Correlation in antonymy

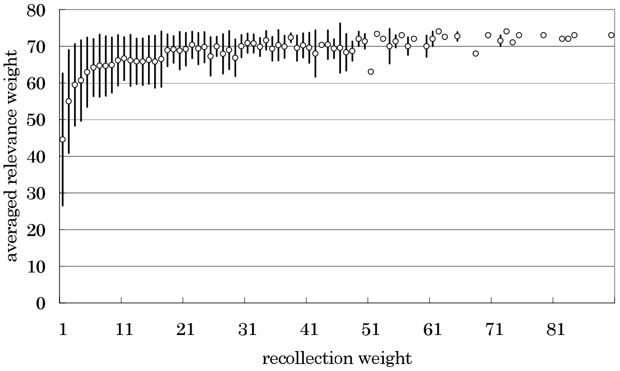

Figure 3. Correlation in associativity

In synonymy and antonymy, the peaks of the distributions are located at one to three subjects and most of the words are judged to be relevant by less than seven subjects (9%). On the contrary, the peak of the distribution in associativity is located at four to seven subjects and there are 54 non-recollected words (5.4%) judged to be relevant by more than half of the subjects.

The above mentioned result can be summarized as follows. There are middle degrees of correlation between recollection weight and relevancy weight in synonymy and antonymy. But in associativity, dependence of the relevancy weight on the recollection weight is little and there are many relevant associative words besides recollected words.

Therefore, when a standard database of only highly related words to a stimulus word is needed in synonymy and antonymy, it is sufficient to measure only recollection weights and to select pairs whose recollection weights are large and whose relevancy are also expected to be large. But when making a standard database of associative words for a stimulus word, it is necessary to measure the relevancy weights between the stimulus word and all the target words which may appear in the simulation to judge word relations.

Construction of standard database of human judgments

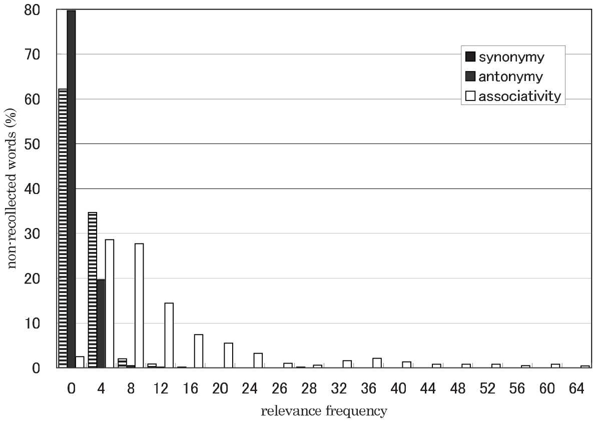

Figure 4. Relevance weight distribution for non-recollected words

From the above mentioned analyses, it is clear that the two experiments are suitable to use for making a standard database of human judgments about the word relations of synonymy, antonymy, and associativity. However, they are not sufficient when making a standard database which contains most of relevant associative words with a stimulus word because non-recollective, but relevant associative words exist.

As an example we made a standard database for a related word retrieval task. During the task, a computer calculates weights between a stimulus word of the 200 Japanese words and each of given target words. Then the computer outputs a list of the target words ranked based on the weights. Usually, a human subject evaluates whether each of the listed words is related to the stimulus word. Therefore, a standard relevancy weight may be suitable for selecting related words to compare with the simulation. For each of 200 stimulus words, we chose target words whose recollection weight was more than two, because we did not measured the relevancy weights of all the recollected words, and whose relevancy weight are more than 38, i.e., over half of the number of all subjects in Experiment 2. As a result, averaged numbers of related target words for each stimulus word were 5.15 in synonymy, 3.49 in antonymy, and 29.85 in associativity.

Comparison of human judgments with a thesaurus

In order to check whether a conventional standard database for the simulation of judging word relationships is effective in a word retrieval task, we used the standard database to evaluate a large-scale thesaurus[7] in which 300,000 Japanese words are categorized into 3,000 categories.

For each of the 200 stimulus words, we chose target words which were in the same category as the stimulus word in random order to make a word list of the related word retrieval. Each related word in the list and the standard related words in the database were compared by using a well-known method for evaluating a retrieval result [4] and the values of precision were calculated in the eleven recall values. (Figure 5). In the three relations, all the precision values are fairly low. Averaged precision values for eleven recall values are 0.03 in synonymy, 0.01 in antonymy and associativity. One of the reasons for these low values is that the evaluated thesaurus contains a huge number of words some of which would be difficult for an average Japanese native speaker.

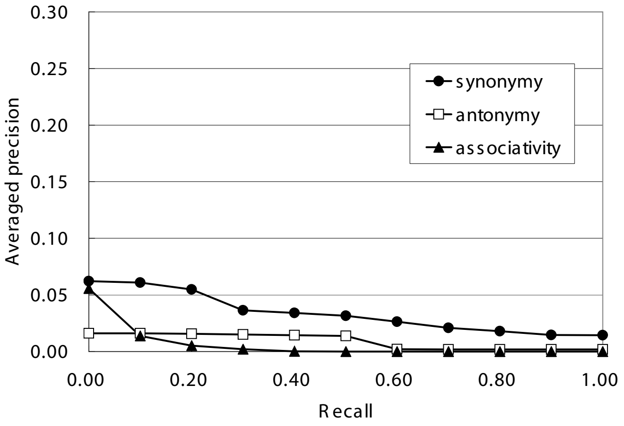

Figure 5. Evaluation of a thesaurus in related word retrieval

Comparison of human judgments with the results of the method of simulation

Figure 6 shows the result of applying the standard database to evaluation of the method for simulating judgment of semantic similarity between words[10]. This method represents a word concept in a vector form whose weights of attributes are decided from word definitions in a machine readable dictionary. The degree of similarity between two words becomes a cosine of an angle between their word vectors. Basically, the closer the definitions of the two words are to each other, the higher the degree of similarity becomes. When this method is applied to simulating similar word retrieval, the degrees of similarity between a given word and all the words which have word vectors are calculated and the words which have high degrees of similarity are output as the result.

Figure 6 shows the precision–recall curves averaged on 200 retrievals in the relations, synonymy, antonymy, or associativity. It shows that this method of simulation is much better than that of applying thesaurus to retrieval for any of the three relations. It also shows that this method simulates judgment of synonymy better than antonymy and associativity because the method was intended to simulate judgment of similarity between words. Therefore, some other revisions to the method may be required in order to be able to use it for the simulation of judging antonymy or associativity.

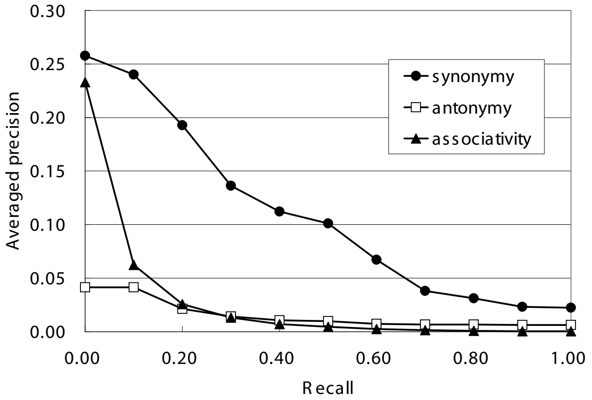

Figure 6. Evaluation of a simulation method

4. Conclusion

In this paper we analyzed correlation of two sorts of weights, recollection and relevancy weight, for the relations, synonymy, antonymy, and associativity based on two human subject experiments. These analyses are important to make a standard database for evaluating the simulation of judging word relations.

The result shows that the two sorts of weight have middle correlation in synonymy and antonymy, but low correlation in associativity. We also built a standard database for a related word retrieval task and found that a large-scale thesaurus is not appropriate to simulate to generate a commonly known related word list. A previously proposed method of simulating judgment of similarity between words based on a machine-readable dictionary was also evaluated and it was found that judging synonymy much better simulated than judging antonymy and associativity.

Acknowledgement

This research was supported in part by the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University(, research project on Concept Bases for Lexical Acquuisition and Intelligently Reasoning with Menings.)

References

[1] S. Amano and T. Kondo. Estimation of mental lexicon size with word familiarity database. In *Proc. of Intl. Conf. on Spoken Language Processing*, volume 5, pages 2119 – 2122, 1998.

[2] S. Amano and T. Kondo. *Goi-Tokusei (Lexical properties of Japanese) Vol. 1*. Sanseido, 1999.

[3] K. Bessho. Text segmentation using word conceptual vectors (in Japanese). *Trans. of IPSJ*, 42(11):2650 – 2662, 2001.

[4] C. Buckley. Smart version 11.0. ftp://ftp.cs.cornell.edu/pub/smart, 1992.

[5] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391 – 407, 1990.

[6] C. Fellbaum, editor. *WordNet : an electronic lexical database*. The MIT Press, 1998.

[7] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Oyama, and Y. Hayashi, editors. *The Semantic System, volume 1 of Goi-Taikei – A Japanese Lexicon*. Iwanami Shoten, 1997.

[8] S. Ikehara, S. Shirai, A. Yokoo, and N. Hiromi. Toward an mt system without pre-editing -effects of new methods in alt-j/e-. In *MT Summit '91*, pages 101–106, 1991.

[9] J. E. D. R. Institute. *EDR Electronic Dictionary Technical Guide*, tr-042 edition, 1993.

[10] K. Kasahara, K. Matsuzawa, and T. Ishikawa. A method for judgment of semantic similarity between daily-used words by using machine readable dictionaries (in Japanese). *Trans. of IPSJ*, 38(7):1272 – 1283, 1997.

[11] H. Moss and L. Older. *Birkbeck Word Association Norms*. Psychology Press, 1996.

[12] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. The university of south florida word association, rhyme, and word fragment norms. In *http://www.usf.edu/FreeAssociation/*, 1998.

[13] H. Schütze. Dimensions of meaning. In *Proceedings of Supercomputing 92*, pages 787–796, 1992.

[14] H. Schütze and J. Pedersen. Information retrieval based on word senses. In *Fourth Annual Symp. on Document Analysis and Information Retrieval*, pages 161–175, 1995.

[15] T. Umemoto. *Renso Kijun Hyou*. University of Tokyo Press, 1969.