International Journal for Housing Science and Its Applications
Publish August 5, 2025. Volume 46, Issue 3 Pages 3736-3745

https://doi.org/10.70517/ijhsa463315

Modeling and deep computational analysis of Japanese lexical semantic relations based on high-dimensional space

Xiaodan Li[1,*]
[1] School of Foreign Languages, Liaodong University, Dandong, Liaoning, 118001, China
Corresponding authors: (e-mail: dandan332210@sina.com).

Abstract Japanese vocabulary corpus is rich and semantically complex, and traditional methods are more limited in dealing with its semantic relations, which makes it difficult to deal with large-scale Japanese vocabulary corpus effectively. In this paper, we first establish the conceptual semantic network of Japanese vocabulary by means of co-occurrence analysis statistics and similarity computation, and construct a vocabulary semantic similarity measure model based on Japanese-specific corpus by extracting the contextual features of Japanese vocabulary. Then, we use machine translation to construct word-vector relations between Japanese and English, and then introduce LSTMs network to learn the sentence sequences co-occurring between Japanese word pairs, so as to complete the modeling of lexical relations between Japanese word pairs. The Japanese vocabulary conceptual semantic network constructed in this paper clearly identifies the center nodes and non-center nodes of the Japanese vocabulary semantic network, and the similarity of Japanese vocabulary pairs under this computation and closer to the fact, compared with the JC algorithm, the algorithm in this paper reduces the gap between the computation results and the manual algorithm to 0.00. In general, the method is conducive to the improvement of the efficiency and the intelligence of the processing of large-scale Japanese vocabulary corpus.

Index Terms Similarity calculation method, Lexical concept semantic network, Contextual context feature extraction, LSTMs network, Skip-gram word vector modeling

I. Introduction

Semantics is the centerpiece of human language communication, which is not only needed for the correct transmission of ideas and social communication, but also urgently needed for the understanding of human natural language by electronic computers [1], [2]. In vocabulary depth knowledge, the semantic relationship between words is what language learners need to focus on. Vocabulary in the learner's mind does not exist in isolation, they are gathered into clusters due to the intrinsic semantic relations, and the clusters of words form a network-like mental lexicon due to various connections between them, and the semantic relations are an important part of the mental lexicon, which is classified by Saussure as the horizontal combinatorial and vertical aggregation relations [3]-[6]. The semantic combination relationship is firstly manifested as the collocation and co-occurrence relationship between words. The collocation of words is subject to syntactic constraints on the one hand, and semantic conditions on the other, and it is because of the strict semantic connection that the combination of words also constitutes a semantic field [7], [8]. Psychologists' studies of human lexical memory have shown that lexical semantic connections form an associative network in memory, which helps to memorize words [9]. Vocabulary longitudinal aggregated semantic relations are crucial for the development of lexical networks because they directly affect the richness of lexical knowledge and the depth of lexical knowledge, and the complexity of vocabulary longitudinal aggregated semantic relations poses a great challenge for language learning, making it necessary to carry out research on the development of lexical semantic relations [10]. Japanese belongs to the adhesive language, and in the Japanese lexical system, Japanese is extremely varied, and there is not only the distinction between spoken and written language. There are also differences between simplified and honorific, ordinary and solemn, male and female, old and young, and people in different professions and positions speak differently, and the form of mixed writing in text writing and the strict honorific system reflect the complex structure of Japanese semantic space, which hinders in-depth exploration of semantic relations [11]-[15].

The aim of this study is to optimize the method of modeling and depth calculation of Japanese lexical semantic relations, and to provide a method for more fine-grained portrayal of Japanese lexical semantic features. Firstly, a Japanese lexical feature mining method based on conceptual semantic network is proposed to establish a Japanese lexical conceptual semantic network. Then the similarity model criterion is used as a control template, and the optimization of the similarity model is completed based on the upper and lower contextual feature vectors of

Japanese words respectively. The word vectors of Japanese vocabulary are obtained using the standard Skip-gram model. Next, machine translation is used to construct Japanese and English word vector relations, and Japanese word vectors are replaced with English word vectors to simplify the process of Japanese word vector computation. Subsequently, the Japanese lexical relations are predicted and the calculation of Japanese lexical similarity is completed to learn Japanese lexical relations with LSTMs network. Finally, the features of Japanese vocabulary concept semantic network are studied and visualized, and the superiority of the improved method in this paper is verified with examples.

II. Modeling and Deep Computing of Japanese Lexical Semantic Relationships

II. A. Constructing a semantic network of Japanese lexical concepts

According to the law of lexical distribution, two words that are semantically similar are substitutable for each other in a given context. Therefore, when words are described in terms of other words with which they frequently co-occur, if the co-occurring vocabulary overlap of two words WI and W2 is higher, the higher their semantic similarity [16] is. According to this principle, the establishment of the conceptual semantic network is mainly divided into the following steps: the lexical extraction stage uses a keyword dictionary to extract words for each transcript in the Japanese vocabulary corpus, and at the same time counts the word frequencies and generates the orthographic file. The co-occurrence analysis stage uses DICE measure, the most commonly used association algorithm in statistical learning, to calculate the association degree between two by two of all the words in the positive ranking file, with the formula:

$$D(A,B)=2\times P(AB)/(P(A)+P(B)) \tag{1}$$

where $P(A)$ and $P(B)$ denote the likelihood of vocabulary A and B appearing alone, respectively, and $P(AB)$ denotes the likelihood of vocabulary A and B appearing at the same time.3 Conceptual Semantic Similarity Calculation Stage. The cosine similarity algorithm commonly used in vector space modeling is used to calculate the semantic similarity between words by taking the first N words with the highest degree of association with the words as their feature vectors, see Eq:

$$Sim(T_i,T_j)=\frac{\sum_{k=1}^{n}t_{ki}t_{kj}}{\sqrt{\sum_{k=1}^{n}t_{ki}^2\sum_{k=1}^{n}t_{kj}^2}} \tag{2}$$

where t_{ki} denotes the correlation degree of the k th feature word of the word i, and $t_{ki}t_{kj}$ denotes the product between the correlation degrees corresponding to the respective feature words of the word i and the word j when their k th feature word is the same. The larger the value of semantic similarity, the higher the semantic similarity between the two words.

After the above steps, the calculation of semantic similarity between all words two by two is completed, and a conceptual semantic network of Japanese words represented by these words is formed. In this network, the nodes in the network are words, and the connecting lines between the words represent the existence of semantic similarity between the two, and the weight of the connecting lines is the semantic similarity between the two.

II. B. Contextual features of vocabulary

The word vector method [17] is a widely used method in statistical linguistics at present, and the related model is characterized by low computational complexity, high sensitivity and easy training. Applying the word vector method to the semantic similarity calculation of words, the above and below neighboring real words of the object words are taken as the contextual feature vectors respectively, and the distance between the lexical contextual feature vectors based on the sample corpus is taken as the semantic similarity measure between the words.

II. C. Semantic similarity computation model

II. C. 1) Model guidelines

In general, lexical semantic similarity is measured using normalized measure with similarity value domain [0, 1]. The lexical semantic similarity calculation model needs to satisfy the following conditions:

(1) The similarity between a lexicon and itself is 1;

(2) If two words are irreplaceable in any context, then their similarity is 0;

(3) the similarity measure is monotonic, i.e., the more semantically similar two vocabularies are, the more similar they are.

For two words S_1 and S_2, we denote their similarity as $Sim(S_1, S_2)$, and any computational model that satisfies the above conditions can be used as a semantic similarity measure.

II. C. 2) Similarity Measurement Based on Contextual Feature Vectors

If the context space of the sample corpus is remembered as $C = \{c_1, c_2, \cdots, c_n\}$, where c_i denotes a real word in the corpus. Remember that the contextual feature vectors of the words S_1 and S_2 are $S_1 = \{s_{11}, s_{12}, \cdots, s_{1n}\}$ and $S_2 = \{s_{21}, s_{22}, \cdots, s_{2n}\}$, where s_{ij} denotes the number of occurrences of the real word c_j in the contextual feature vector of the i th word.

Against the similarity modeling criterion, based on the contextual feature vectors of the vocabulary, we construct the similarity computation model based on the contextual feature vectors as follows:

$$\begin{aligned} Sim(S_1, S_2) &= \sqrt{\frac{2\sum_{j=1}^{n} s_{1j} \cdot s_{2j}}{\sum_{j=1}^{n} s_{ij}^2 + \sum_{j=1}^{n} s_{2j}^2}} \\ &= \sqrt{\frac{2S_1S_2^T}{S_1S_1^T + S_2S_2^T}} \end{aligned} \tag{3}$$

Easy to know $Sim(S_1, S_2) \in [0,1], Sim(S_1, S_1) = 1$, if there is a vocabulary $S_3 = \{s_{31}, s_{32}, \cdots, s_{3n}\}$, then:

$$\begin{aligned} & Sim^2(S_1, S_2) - Sim^2(S_3, S_2) \\ &= \left(\frac{2S_1}{S_1S_1^T + S_2S_2^T} - \frac{2S_3}{S_3S_3^T + S_2S_2^T}\right)S_2^T \\ &= \frac{2(S_2S_2^T - S_1S_3^T)(S_1 - S_3)S_2^T}{(S_1S_1^T + S_2S_2^T)(S_3S_3^T + S_2S_2^T)} \end{aligned} \tag{4}$$

Assuming that $s_{31} \neq s_{11}, s_{32} = s_{12}, s_{33} = s_{13}, \cdots, s_{3n} = s_{1n}$ (the case of $s_{3i} \neq s_{1i}$) Similarly), we find from the above equation that when $|s_{31} - s_{21}| < |s_{11} - s_{21}|$, there is $Sim(S_3, S_2) > Sim(S_1, S_2)$; and when $|s_{31} - s_{21}| > |s_{11} - s_{21}|$, then we have $Sim(S_3, S_2) < Sim(S_1, S_2)$. That is, when the projection of the lexicon S_3 in the sample context space is closer to S_2 compared to the lexicon S_1, its semantic similarity is higher. Thereby, the model constructed above meets the general criterion of lexical semantic similarity measure.

II. C. 3) Model Optimization for Contextual Contexts

For the linguistic characteristics of Japanese, the similarity model is optimized as follows by assigning different weights to the upper and lower context spaces respectively:

$$\begin{aligned} Sim(S_1, S_2) = \alpha & \sqrt{\frac{2S_{U,1}S_{U,2}^T}{S_{U,1}S_{U,1}^T + S_{U,2}S_{U,2}^T}} \\ +\beta & \sqrt{\frac{2S_{D,1}S_{D,2}^T}{S_{D,1}S_{D,1}^T + S_{D,2}S_{D,2}^T}} \end{aligned} \tag{5}$$

where $S_{u,i}, S_{D,i}$ denote the upper and lower contextual feature vectors of the vocabulary S_i respectively, and (α, β) is the weight vector, taking into account the characteristics of the “semantic backwardness” of the Japanese language, in general, we configure $\beta > \alpha$.

II. D.Lexical Semantic Similarity Calculation Based on Improved Algorithm

II. D. 1) Standard Skip-gram Word Vector Modeling

The core idea of the Skip-gram model [18] is to build a three-layer neural language model of input-mapping-output layer, which utilizes each current word w_t as the input to a linear Logistic classifier in the projection layer to predict words w_t within a certain range before and after $w_{t-T}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+T}$.

Let $W^{(1)}$ denote the matrix of word vectors, where each word in the word list can be mapped into a continuous vector of values by looking up the matrix, and $W^{(2)}$ denote the matrix formed by the surrounding words in the window. Given the training data $w_1, w_2, ..., w_n$ sequence, then the model actually aims to maximize the function:

$$Q = \frac{1}{N}\sum_{n=1}^{N}\sum_{-c\leq j\leq c, j\neq 0} \log p(w_{n+j} \mid w_n) \tag{6}$$

where N represents the number of words, c is the context window size, w_n represents the center word, w_{n+j} is the neighboring word of the center word, and the conditional probability $p(w_{n+j} \mid w_n)$ can be expressed as:

$$p(w_{n+j} \mid w_n) = \frac{\exp(w_{n+j}^{(2)} \cdot w_n^{(1)})}{\sum_{k=1}^{V}\exp(w_k^{(2)} \cdot w_n^{(1)})} \tag{7}$$

where $w_n^{(1)}$ and $w_k^{(2)}$ denote the row vectors in the matrices $W^{(1)}$ and $W^{(2)}$.

Therefore, as long as a large-scale contextual corpus is found, a distributed representation of the word vector space can be trained based on the above model, and then, the distances of the semantic space vectors are utilized to compute to get the final word semantic similarity scores.

II. D. 2) Improved methods based on machine translation

Japanese corpus training of word vectors encounters the problem of word separation error, which may affect the training effect of the corpus. In contrast, English vocabulary has natural word separators and does not suffer from such problems as word separation errors.

Therefore, based on the above facts, we ask the question: can we utilize the English public word vector model trained on massive data to improve the performance of Japanese word similarity computation?

Under this assumption, the problem we need to solve is how to evaluate whether the machine translation is completely translated correctly. Specifically, after Japanese words are translated into English, on the one hand, the spelling of the English translation is required to be free of errors, and on the other hand, the semantics of the translated words are required to be unchanged. In this paper, two strict constraints are set to ensure the translation quality:

(1) An English translation is free of spelling errors after spell-checking.

(2) A Japanese word is translated into only one English word.

The specific steps to achieve this are:

(1) Construct a query Query using each word and call Google Translation API to translate each set of Japanese words into English words corresponding to them respectively.

(2) Check and mark two word pairs whose translations are both 1 word in length.

(3) Detect the word pairs marked in step (2) using the word spelling error checker, and eliminate the word pairs with spelling errors. In this paper, the algorithm provided by PyEnchant toolkit is used for word spelling detection.

(4) The remaining word pairs are taken out of the English word vector model for similarity computation, and other word pairs are still computed using Japanese word vectors.

II. D. 3) Improved method based on LSTMs

This paper discusses how to train a Skip-gram model using a large real corpus of lexical occurrences to obtain a distributed word vector space that can represent semantics. On the basis of the word vectors obtained from the training of standard Skip-gram models, this section further discusses how to construct a neural language model using LSTMs models to make predictions on similarity scores.

Our basic assumption is that in addition to the contextual environment of words that can reflect their semantics, for the lexical similarity evaluation task, the contextual environment in which two words in a word pair co-occur can also strengthen the semantic relationship between them. Therefore, in this paper, we want to construct such a process for LSTMs model to learn sentence sequences: in the training process, the inputs are sentence sequences co-occurring in word pairs and integer labels $i = 1, \ldots, 10$ after rounding the similarity scores, and train the LSTMs model accordingly; in the testing process, the inputs are sentence sequences co-occurring in word pairs and the trained LSTMs model, and the outputs are the semantic similarity scores of the final predicted word pairs.

Assuming that the sentence S in which the word pair (w_1, w_2) is located is to be learned, it is necessary to consider how to construct the input vectors and how to update each of the weights in the model, respectively.

First, we construct a $4n(n=150)$ -dimensional word vector with distance information: where the first $2n$ dimensions are the pre-trained word vectors of the Skip-gram, the next n dimensions are populated with the distances of the current word from w_1, and the last n dimensions are populated with the distances of the current word from w_2, and the distance value is positive if the current word is on the right-hand side of the target word. If the current word is on the right side of the target word then the distance value is positive, if the current word is on the left side of the target word then the distance value is negative, and if the target word is the current word itself then the distance value is zero.

In the following paper, we will further illustrate how the memory neuron updates the parameters at each moment with the formula, so that x_i is the input of the memory neuron at the t th moment, $W_i, W_f, W_c, W_o, U_i, U_f, U_c, U_o, V_o$ are the weight matrices and b_i, b_f, b_c, b_o are the bias vectors, respectively.

First, the input gate activation value i_t and the candidate value $\tilde{C}_t$ for the state of the memory neuron are computed at the t th moment:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{8}$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{9}$$

Next, calculate the activation value f_t of the forget gate at the t th moment:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{10}$$

Given the input gate activation value i_t, the forget gate activation value f_t, and a candidate value for the state of the memory neuron $\tilde{C}_t$, the new memory neuron state at the t th moment can be calculated as:

$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1} \tag{11}$$

Then, using the newly memorized nerve cell states, their output gate activation values and final outputs can be computed, respectively, and are expressed as Eq:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{12}$$

$$h_t = o_t * \tanh(C_t) \tag{13}$$

The sequence of sentences $x_0, x_1, ..., x_n$ is input into the LSTMs layer will produce a new sequence $h_0, h_1, ..., h_n$, subsequently, this sequence is averaged over all time steps through the Mean Pooling layer, and then enters the Logistic Regression layer to output the probability that each category label is i as p_i, and ultimately, the similarity scores between word pairs can be obtained as:

$$auto\_score = \sum_{i=1}^{10} i * p_i \tag{14}$$

The specific steps to realize this are:

(1) Crawl the text fragments co-occurring in word pairs. In this paper, we construct a query Query with words w_1 and w_2 to send a query request to Baidu API, and crawl the collection of retrieval result pages to get the snapshots of web pages containing both words w_1 and w_2.

(2) Preprocessing to get word co-occurring sentences. In order to clean the original data to obtain a training corpus more favorable for semantic extraction, this paper carries out preprocessing work on the original web page snapshot fragments.

(3) Model training and testing. Build the LSTMs network framework, and randomly sample the corpus consisting of word pairs and sentences into five parts, one of which is used as the test corpus, and the other four are combined as the training corpus, which constitutes five experimental groups, i.e., five-fold crossover experiments are conducted.

III. Characterization and Visualization of Conceptual Semantic Networks of Japanese Vocabulary

III. A. Descriptive features of Japanese lexical semantic networks

A total of 200 valid vocabularies were collected in this study. Therefore, the size of the constructed lexical semantic network is 200; however, the frequency of output varies among different vocational words. The 10 words with the highest output frequency are as follows: transport, car, cycle, road, driver, bicycle, men, woman, singer, and worker, and they are numbered 1-10 respectively. Table 1 shows the connections between the 10 words with the highest frequency of output, and the values on the diagonal in the table are the frequency of that word's output. The average path length of the network is 1.844, which means that each node can be associated with all other nodes after an average of 1844 edges. The average density of the network is 0.4251, which means that only 42.51% of all possible node connections actually occur, so there is a lot of room for improvement in the density of this network.

Table 1: The connection between the words of the first 10 of the output frequency

	1	2	3	4	5	6	7	8	9	10
1	15	20	15	17	20	15	14	18	20	16
2	20	16	13	20	13	15	11	9	21	23
3	15	13	19	24	23	21	25	18	16	21
4	17	20	24	15	14	21	22	16	14	21
5	20	13	23	14	21	22	25	21	15	19
6	15	15	21	21	22	14	16	14	9	8
7	14	11	25	22	25	16	18	19	15	16
8	18	9	18	16	21	14	19	22	23	20
9	20	21	16	14	15	9	15	23	11	13
10	16	23	21	21	19	8	16	20	13	20

III. B. Japanese lexical semantic network partition density distribution

In order to find out the central node by density comparison, it is necessary to minimize the bias caused by the huge difference between the node degrees, so we first normalize the word-word matrix, and then run the Ucinet software for Concor analysis of vocational vocabulary semantic network, and the results of semantic vocabulary module analysis are shown in Table 2.

Table 2 is the density table of the Japanese vocabulary network, which clearly shows the distribution of connections within and between modules. The values on the diagonal line are the densities within each module, e.g., the internal densities of Blocks 1-4 are 0.0366, 0.0190, 0.0142, 0.0102, respectively, decreasing in order, and the internal densities of Blocks 5-8 are 0.0036, 0.0016, 0.0246 , 0.0335, decreasing then increasing, the change rule is obvious. From the change of internal density, it can be seen that Blocks1-4 and Blocks5-8 form two large modules respectively. And from the connection density between modules, it can be seen that not only the internal density of modules in Blocks1-4 is large, but also the connection density between modules is larger, especially the connection density between Block1 and Block2 and all other modules is above 0.0133, which belongs to the two most central modules. If we divide all nodes into only two major modules, the result is that Blocks1-4 constitute the central module while Blocks5-8 constitute the non-central module. Accordingly, the words in Blocks1-4 are classified as network center node words.

Table 2: The semantic density distribution after the partition

	Block1	Block2	Block3	Block4	Block5	Block6	Block7	Block8
Block1	0.0366	0.0324	0.0300	0.0347	0.0214	0.0339	0.015	0.0132
Block2	0.0165	0.0190	0.0152	0.0136	0.0199	0.0165	0.0184	0.0124
Block3	0.0095	0.0088	0.0142	0.0077	0.0173	0.0096	0.0029	0.0103
Block4	0.0097	0.0078	0.0079	0.0102	0.0034	0.0095	0.0022	0.0093
Block5	0.0056	0.0018	0.0023	0.0006	0.0036	0.0013	0.0026	0.0015
Block6	0.0014	0.0019	0.001	0.0022	0.0008	0.0016	0.0016	0.0014
Block7	0.0045	0.0016	0.0018	0.0051	0.0017	0.0062	0.0246	0.0021
Block8	0.0032	0.0016	0.0021	0.0023	0.0015	0.0008	0.0036	0.0335

III. C. Visualization of Japanese lexical semantic networks

Japanese lexical semantic networks are constructed in two steps; first, Japanese words are organized into the basic skeleton of the network using the superordinate and infrasegmental relations, and then each pair of words is examined in turn to see if there are any other connections between the words, and if so, new relational connectives are added between the pairs of words. Figure 1 shows the relational network diagram of some Japanese words constructed in this paper.

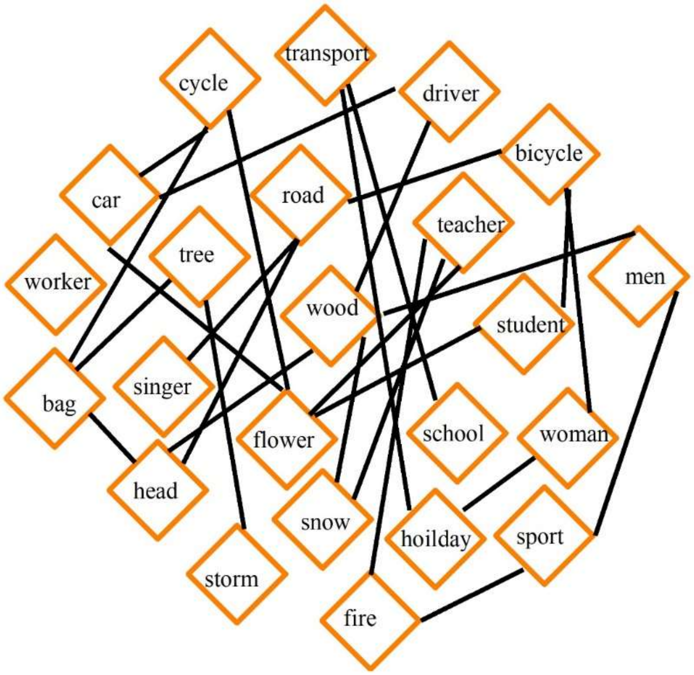

Figure 1: The Japanese vocabulary network is part of the structure

Combined with the practical application requirements of image tag sorting, we downloaded 1,000 images with the tags “traffic” or “vehicle” from the image sharing website Flickr, and after preprocessing, these 1,000 images contained 2,000 unduplicated tags. After preprocessing, these 1000 images contain 2000 non-repeating tags, and the tags with the top 300 occurrences, including traffic, vehicle, car, people, street, and so on, are numbered from 1 to 300 as the commonly used words for constructing the Japanese vocabulary domain knowledge network. In order to intuitively evaluate the calculation results of similarity, the WP algorithm based on path length, the JC algorithm based on information content, and the algorithm in this paper are used to calculate the semantic similarity of some Japanese words, respectively. Figure 2 shows the similarity change curves after normalization for Japanese words numbered 1-20.

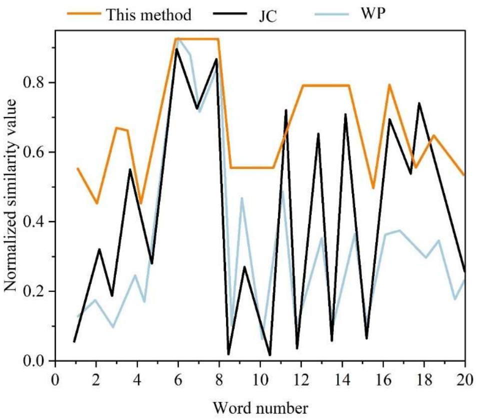

Figure 2: Some semantic similarity of some Japanese vocabulary

The similarity curves of the WP algorithm and the JC algorithm are basically in the same trend except for the difference in the bias amount. In this paper, the similarity of domain knowledge is strengthened by the algorithm due to the introduction of multiple lexical relations, which is overall higher than the WP and JC algorithms. Compared with the WP and JC algorithms in which some vocabulary similarity values are on the small side, this paper's algorithm adopts relations such as subject-event and place-event to overcome the natural “weak similarity” between abstract things and concrete things, so that the vocabulary similarity values on the small side are significantly increased.

Finkelstein presented a semantic similarity test set containing 350 pairs of words. Each pair of words in the test set was carefully selected by experts, covering types ranging from "highly semantically relevant" to "semantically irrelevant". In order to obtain a real human evaluation, Fifty subjects were invited to independently rate the "meaning similarity" of these 350 pairs of words, with the scores ranging from 0.0 to 4.0. The average of the subjects' scores is the true value of the test set. Pearson's correlation is a common criterion for evaluating the goodness of a vocabulary similarity algorithm, which reflects the extent to which the similarity value obtained by the algorithm conforms to the results of manual judgment in the Finkelstein test set, and the higher the correlation, the better the algorithm is. Fifty pairs of Japanese words were selected from the Finkelstein test set, and JC algorithm and this paper's algorithm were used to calculate the similarity, and the Pearson correlation and partial calculation results of each algorithm are shown in Table 3.

Overall, the calculation results of this paper's algorithm are closer to the real value and have higher Pearson correlation values. The Pearson correlation values are shown in Table 3, and in the similarity calculation of some word pairs, the results of this paper's algorithm are significantly better than the JC algorithm. As for steering and vehicle, this paper's algorithm obtains a similarity value that is more in line with the manual judgment, and the difference between the similarity value and the manual judgment is 0.00, which is reduced by 0.32 compared to the gap between the JC algorithm and the manual algorithm.

Table 3: The similarity of some words to different algorithms

Word pair	True value	JC algorithm	This algorithm
Car-driveing	0.80	0.37	0.67
Car-journey	0.67	0.12	0.62
Cushion-automobile	0.30	0.45	0.17
Automobile-car	0.96	0.98	0.96
Vehicle- underground	0.88	0.06	0.88
Steering-vehicle	0.67	0.35	0.67
Track-train	0.79	0.57	0.64
Street-highway	0.81	0.60	0.68

IV. Example of Semantic Relevance Calculation

IV. A. Experimental environment and data set

MongoDB is a document-oriented non-relational database based on distributed file storage, which is more suitable for large-scale data storage than traditional relational databases. In this paper, we build localized dataset by MongoDB to invoke queries.

IV. B. Experimental cases and operational results

Table 4 shows the 20 sets of Japanese word pairs used as experimental cases. Compared with the JC calculation method, the values derived from the calculation method proposed in this paper have better discriminability, as analyzed below:

(1) Observing "man" and "father," the result values of the three methods compared in the JC calculation are all 1. However, "father" must be a "man," while a "man" is not necessarily a "father." Therefore, the judgment result is 1, which means that semantic equivalence is not entirely reasonable and deviates from common sense.

(2) Observing the experimental results of pairs such as "Japan" and "America," "Japanese" and "Americans," "Cambridge University" and "Harvard University," the main attributes are synonymous, but the semantics are not entirely the same. A result value of 1 does not align with the facts. Analyzing all data in the table, it shows that the trend of the experimental results is stabilizing, but there are also occurrences within the result value data set.

Table 4: Comparisons of the calculating results of semantic relatedness

Vocabulary 1	Vocabulary 2	JC	Method of this article (relationship value)
Men	Woman	0.97	0.80
Men	Father	1.00	0.37
Men	Mother	0.94	0.84
Men	Monk	0.86	0.71
Men	Manager	0.65	0.65
Men	Radio	0.02	0.21
Men	Pleasure	0.13	0.27
Men	Carp	0.38	0.43
Men	Apple	0.25	0.22
Men	Work	0.10	0.17
Men	Liability	0.08	0.26
Vehicle	Cloudburst	0.13	0.23
Vehicle	Silkworm	0.20	0.33
Cold food	Garbage	0.23	0.21
Japan	The United States	1.00	0.70
Japanese	American	1.00	0.81
Japanese	Roosevelt	0.73	0.63
Japanese army	U.S. military	0.84	0.72
Democratic Party	Democrats	0.85	0.87
Cambridge university	Harvard	1.00	0.96

The test results obtained by applying the improved Japanese lexical semantic similarity calculation method based on this paper to this experimental case are shown in Table 5. The calculated result is less than the threshold value of 0.83, which means that the detected relationship is semantic relatedness. For example, “man” and “responsibility” have a correlation of 0.33 and are iterative semantic relations.

Combining the results in Table 4 and Table 5, it can be seen that the semantic relationship detected by this method is richer and more suitable for supplementing the semantic relationship between two words in terms of semantic relatedness, and the data volume included in this method is larger, which can expand the word retrieval scope of Japanese semantic detection.

Table 5: Semantic correlation calculation results

Vocabulary 1	Vocabulary 2	Semantic relation	This method
Men	Woman	Iteration	0.83
Men	Father	Iteration	0.56
Men	Mother	Iteration	0.33
Men	Monk	Iteration	0.56
Men	Manager	Iteration	0.33
Men	Radio	-	0
Men	Pleasure	Iteration	0.56
Men	Carp	-	0
Men	Apple	Iteration	0.33
Men	Work	Iteration	0.56
Men	Liability	Iteration	0.33
Vehicle	Cloudburst	-	0
Vehicle	Silkworm	-	0
Cold food	Garbage	-	0
Japan	The United States	-	0
Japanese	American	Iteration	0.24
Japanese	Roosevelt	-	0
Japanese army	U.S. military	-	0
Democratic Party	Democrats	-	0
Cambridge university	Harvard	-	0

V. Conclusion

In this paper, we have captured the deep semantic relationships that exist in Japanese vocabulary by constructing a Japanese vocabulary conceptual semantic network and improving the Japanese vocabulary semantic similarity calculation model.

All the Japanese lexical conceptual semantic network nodes in this study are only divided into Blocks 1-4 constituting the center module and Blocks 5-8 constituting the non-center module. The Japanese words in the center module are the core part of the Japanese lexical semantic network and are the necessary foundation for the development of the Japanese lexical semantic network. The word pairs in the non-central module enrich the entire conceptual semantic network of Japanese vocabulary. The study clearly delineates the central nodes of the Japanese lexical semantic network and provides a methodology for the efficient mode of operation of the model in specific Japanese semantic categories.

The Japanese lexical semantic similarity calculation method in this paper clearly defines the Japanese lexical concepts and detects not only the semantic similarity between Japanese word pairs, but also the semantic relatedness between Japanese word pairs. Compared with the comparison algorithm JC, the similarity of Japanese word pairs calculated by this paper's algorithm is more consistent with the factual situation.

All the above results show that the method of this paper can further optimize the modeling of Japanese lexical semantic relations and deeply reveal the semantic relations existing between words. It provides new tools and methods for Japanese semantic research.

Funding

This work was supported by China Scholarship Council under “The State Scholarship Fund to pursue study” (File No: [2021]340).

References

[1] Beltrama, A. (2020). Social meaning in semantics and pragmatics. Language and Linguistics Compass, 14(9), e12398.

[2] Mokos, K., Nestoridis, T., Katsaros, P., & Bassiliades, N. (2022). Semantic modeling and analysis of natural language system requirements. IEEE Access, 10, 84094-84119.

[3] Geeraerts, D. (2017). Lexical semantics. Dalam Oxford Research Encyclopedia of Linguistics, oleh Dirk Geeraerts. Oxford: Oxford University Press. https://doi. org/10.1093/acrefore/9780199384655.013, 29.

[4] Washio, K., Sekine, S., & Kato, T. (2019, November). Bridging the defined and the defining: Exploiting implicit lexical semantic relations in definition modeling. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3521-3527).

[5] Zhao, Y., Yin, J., Zhang, J., & Wu, L. (2023). Identifying the driving factors of word co-occurrence: a perspective of semantic relations. Scientometrics, 128(12), 6471-6494.

[6] Gruenenfelder, T. M. (2020). The Representation of Coordinate Relations in Lexical Semantic Memory. Frontiers in Psychology, 11, 98.

[7] Phoocharoensil, S. (2021). Multiword Units and Synonymy: Interface between Collocations, Colligations, and Semantic Prosody. GEMA Online Journal of Language Studies, 21(2).

[8] Paour Abed, M. J., Fare Shirazi, S. H., & Heydari, A. (2022). A Study of Semantics with the Collocations of the word. Literary Studies of Islamic texts, 6(24), 61-81.

[9] Lau, M. C., Goh, W. D., & Yap, M. J. (2018). An item-level analysis of lexical-semantic effects in free recall and recognition memory using the megastudy approach. Quarterly Journal of Experimental Psychology, 71(10), 2207-2222.

[10] Heylen, K., & Ruette, T. (2013). Degrees of semantic control in measuring aggregated lexical distances. Approaches to measuring linguistic differences, 361-382.

[11] Lucas, M. (2020). Agglutination, Loanwords, and Japanese Morpholexical Categoryhood: Cross-linguistic Factors in the Grammatical Handling of Verbs and Adjectives in Written English. Linguistics Journal, 14(1).

[12] Jarosz, A. (2017). Japonic languages: an overview. Silva Iaponicarum.

[13] Qian, C. (2023). Research on the impact of honorifics in Japanese on social relationships. Academic Journal of Humanities & Social Sciences, 6(24), 82-87.

[14] Inoue, T., Georgiou, G. K., Imanaka, H., Oshiro, T., Kitamura, H., Maekawa, H., & Parrila, R. (2019). Cross-script transfer of word reading fluency in a mixed writing system: Evidence from a longitudinal study in Japanese. Applied psycholinguistics, 40(2), 235-251.

[15] Diegoli, E., & Öhman, E. (2024). Contrasting the semantic space of ‘shame’and ‘guilt’in English and Japanese. Language and Cognition, 16(4), 1296-1318.

[16] Jie Guo,Shujie Lan,Bin Song & Mengying Wang. (2025). Video-text retrieval based on multi-grained hierarchical aggregation and semantic similarity optimization. Neurocomputing,638,130152-130152.

[17] Sheng Hua Xiong,Zhi Hong Wang,Zhen Song Chen,Gang Li & Hao Zhang. (2025). Text classification of public online messages in civil aviation: A N-BM25 weighted word vectors method. Information Sciences,704,121956-121956.

[18] Enes Celik & Sevinc Ilhan Omurca. (2024). Skip-Gram and Transformer Model for Session-Based Recommendation. Applied Sciences,14(14),6353-6353.