Advance Publication

IEEJ Journal of Industry Applications
J-STAGE Advance published date : Feb. 17, 2023
DOI : 10.1541/ieejjia.22006784

Paper

Topic Selection Using Conceptual Distance: How to Select Topics that are Interesting but Unfamiliar to Users

Yuya Sakai*[a] Non-member, Mitsuharu Matsumoto* Non-member

(Manuscript received Jan. 00, 20XX, revised May 00, 20XX)

In this study, we established a topic selection method that recommends topics that are interesting and unfamiliar to users. To achieve this aim, we used conceptual distance to identify topics that were unfamiliar to users and improved the accuracy of this method by removing conceptually similar words. Many words used in conversations are excluded in the dictionaries and thesauruses. Thus, we developed a model for conceptual distance measurement using machine learning to measure conceptual distances even for such words. By conducting the subject experiments, we confirmed that the established system recommends topics a user is interested in but unfamiliar with compared with the baseline method developed in a previous research.

Keywords : Topic selection, dialog system, conceptual distance, thesaurus dictionary

1. Introduction

Recently, with the spread of smartphones and tablets, dialogue systems, such as Google Assistant (Google) [1,2] and Siri (Apple) [3], have become popular. However, these dialogue systems are not yet pervasive. Most of these dialogue systems are question-answering systems. A question-answering system is a system in which a user inputs a question into the dialogue system and receives an answer. The ability of the question-answering system reduces its usage. The study on dialogue systems [4,5] classifies them into task-oriented and non-task-oriented dialogue systems. A task-oriented dialogue system interacts with people for a specific purpose. Its examples include tourist information [6], virtual coaching [7], restaurant searches [8], and taxi booking [9]. Conversely, a non-task-oriented dialogue system is one with no purpose other than chatting. It is also known as a chatbot [10–13]. Chat is important for interacting with people [14]. Miyashita et al. reported on the behavior of robots in shopping malls [15]. Their study showed that in addition to introducing stores and products, chatting, such as self-disclosure of the robot, affected purchasing behavior. For machines to have more natural conversations with humans, the system must actively present topics to users.

A topic selection survey [16] showed that the earliest studies on topic selection started in the 1990s. Regarding topic selection, Mikami et al. reported a method for changing topics considering the irrelevance of words [17]. This approach is useful for changing topics in the middle of a conversation, but it cannot be used to provide a topic in a conversation system. Other studies have estimated topics in chats, but only a few focused on how to choose a topic at the beginning of a chat. Hence, we examined how to select the topic needed at the beginning of a chat. For the system to select an adequate topic, it must keep up with new words daily. It is also preferable that the selected topic is one the user is interested in but slightly familiar with. To address this issue, we used Twitter as an information source. Several studies have estimated user interest related to Twitter-based topic selection. For instance, Kondo et al. reported on a method for estimating users' interests from Twitter using Latent Dirichlet Allocation (LDA) [18]. They applied the LDA method to the Twitter data of users and their followers to estimate users' interests. This method can extract a user's interest from the tweets of followers. However, users are accustomed to regularly seeing their followers' tweets on their timelines, making it difficult to provide new topics using this method. Consequently, we think it is difficult for users to find unfamiliar topics from followers' tweets. The method of estimating user interest from Twitter relates to a topic selection study using Twitter. For instance, Jilin et al. tried effective advertising using Twitter [19]. Their method estimated the user's interests but did not consider whether the user was familiar with the selected tweet.

In this paper, user interest is estimated based on a Twitter keyword search. Since the topic candidates obtainable by keyword searches are irrelevant to the user's tweet, the user may be partially familiar with them. Therefore, a user can get a topic with which they are unfamiliar. In a previous study, we confirmed that the results obtained using keyword searches include many words the user is interested in [20]. Although past studies could estimate user interest, the selected words were sometimes unfamiliar to the user. To improve the past study, we utilized the conceptual distance between words in a thesaurus dictionary. We regarded words that were far from the known words by the user as topical word candidates [21]. Since not all words that can be topical words are in the thesaurus dictionary, we also build a conceptual distance measurement model that can handle all words using machine learning.

This paper is based on Reference (21), which published in the the IPEC-HIMEJI2022 Conference (2022) ©2022 IEEJ.

a)Correspondence to: Mitsuharu Matsumoto. Email: mitsuharu.matsumoto@ieee.org
* University of Electro-Communications, 1-5-1, Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan

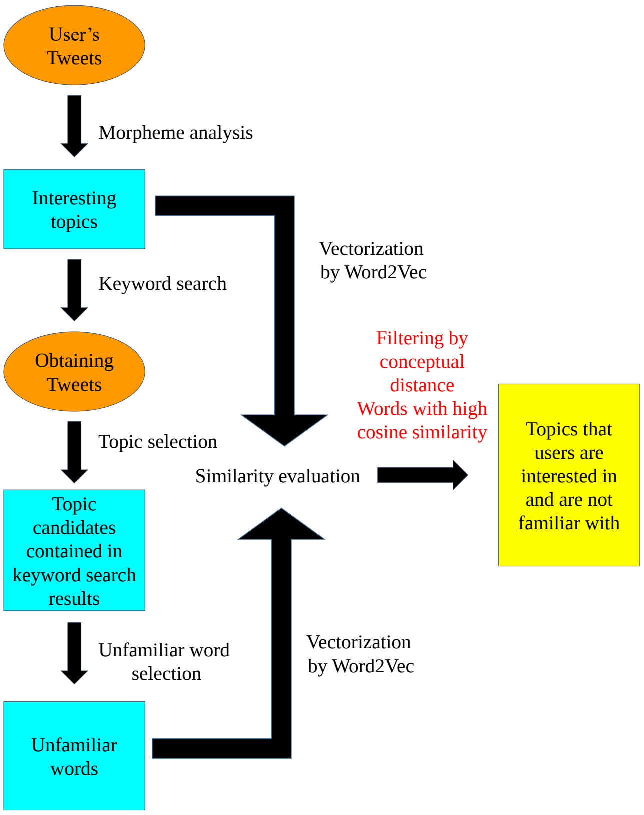
User's Tweets
Morpheme analysis
Interesting topics
Keyword search
Obtaining Tweets
Topic selection
Topic candidates contained in keyword search results
Unfamiliar word selection
Unfamiliar words
Vectorization by Word2Vec
Vectorization by Word2Vec
Filtering by conceptual distance
Words with high cosine similarity
Similarity evaluation
Topics that users are interested in and are not familiar with

Fig. 1. Flow of the proposed method

This study focused on the shortcomings of the previous study [20] and improved it. We improved the previous research in two aspects.

1. Introducing the inverse document frequency (IDF) method from the morphological analysis.
2. Introducing conceptual distance in addition to cosine similarity.

The rest of the paper is summarized as follows: In the next section, we surveyed related works on dialog systems and topic selection using Twitter. In Section 3, we describe the proposed method and its features to extract the topics that users are interested in but unfamiliar with. Section 4 presents the experimental setup and some results on selecting topics using the proposed method. The conclusion and future work are described in Section 5.

2. Related works

2.1 Related works on topic selection from Twitter

There are many approaches to topic selection using Twitter [22]. Many techniques for detecting topics from Twitter are based on known data mining techniques used in various fields. To derive topics from documents, many methods have been reported.

Latent semantic analysis (LSA) is one of the earliest approaches [23]. The LSA uses the singular value decomposition (SVD) method to decompose a matrix into subrepresentations. Hofmann proposed probabilistic latent semantic analysis (PLSA) to handle different meanings and types of words [24]. However, it is an improved version of the LSA. Some approaches use non-negative matrix factorization (NMF) [25]. NMF is a method of decomposing a given matrix into low-dimensional matrix products.

LDA is a generative probabilistic model for document collections [26]. The original LDA model was based on the variational method and the expectation maximization (EM) algorithm for Bayes parameter approximation.

These techniques focus on revealing the semantic relationships between words in a document and have been applied to relatively long texts, such as emails [27,28,29], academic papers [30,31], and web pages [32,33].

However, it is relatively difficult to derive topics from Twitter compared to long documents. For example, the contents are very sparse since tweets are very short and include many incorrect words. A topic can quickly grow, decay, or even merge with another topic. Tweets are fast, and topics change frequently. Despite its difficulty, many studies have applied existing methods. For example, Prier et al. applied the LDA method to a Twitter dataset to derive topics [34]. Kireyev et al. proposed a modified LDA method to identify topics from tweets [35]. Zhang et al. proposed an approach to obtaining hot topics from Twitter [36]. Furthermore, Weng et al. merged tweets into a single document and applied the original LDA method [37]. However, merging all tweets into a single document makes guessing the topic of each tweet difficult.

Some approaches use external document data as an additional dataset to solve the sparsity problem of Twitter. For example, Phan et al. used an external dataset as an additional dataset [37,38]. Hu et al. utilized Wikipedia and WordNet as multiple semantic knowledge [40]. Although using an external resource as a dataset appears promising, the external data does not necessarily contain all the words on Twitter.

2.2 Related works to measure the similarity

The similarity between words or sentences must be measured in the dialogue system. According to a survey on dialogue systems [41], there are three types of measures.

One approach is surface form similarity. Levenshtein distance, METEOR [42], and TF-IDF retrieval models [43,44] are examples of this category. Another approach is multiclass classification, in which the problem is regarded as the problem of multiclass classification [45]. The other approach is the neural network–based approach. Lowe et al. proposed a dual-encoder architecture [46].

Our previous approach measured the distance between words using cosine similarity. However, in this study, we introduced TF-IDF models to improve the performance of the proposed method. We also introduced the conceptual distance to select semantically distant words.

3. Proposed method

3.1 Outline of the proposed method

This section describes the outline of the proposed method. Figure 1 shows the outline of the proposed method. This research aims to build a system that provides topics that users are interested in but are rather unfamiliar with. The system collects data from users' tweets to find out what they are interested in. As the content described in the user's tweet is known to the user, the system performs a keyword search on Twitter using the acquired data and extracts related words. The system analyzes the extracted words using Word2vec and provides words that meet the following three conditions as topic candidates:

1. Words not included in the data obtained from the user.
2. Words that are not too small in concept distance from words that the user knows.

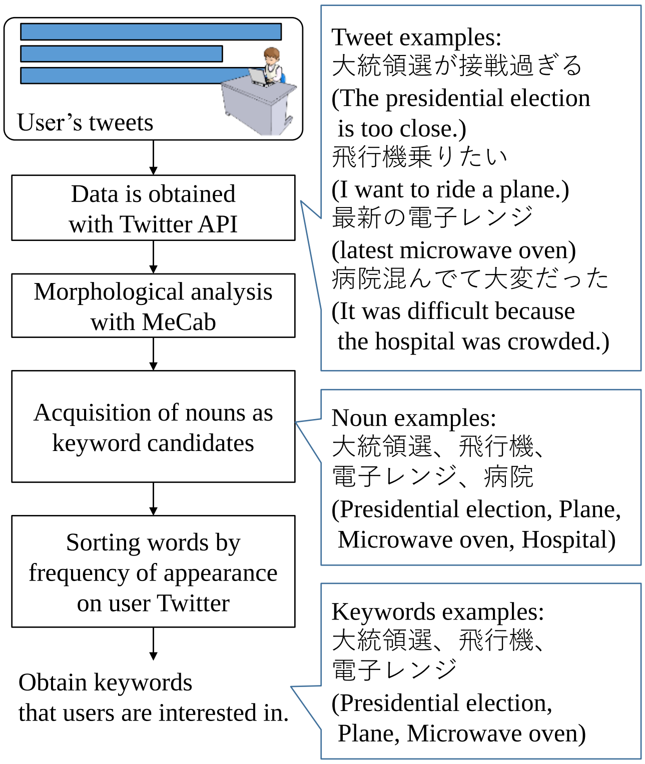
User's tweets
Data is obtained with Twitter API
Morphological analysis with MeCab
Acquisition of nouns as keyword candidates
Sorting words by frequency of appearance on user Twitter
Obtain keywords that users are interested in.
Tweet examples:
大統領選が接戦過ぎる
(The presidential election is too close.)
飛行機乗りたい
(I want to ride a plane.)
最新の電子レンジ
(latest microwave oven)
病院混んでて大変だった
(It was difficult because the hospital was crowded.)
Noun examples:
大統領選、飛行機、電子レンジ、病院
(Presidential election, Plane, Microwave oven, Hospital)
Keywords examples:
大統領選、飛行機、電子レンジ
(Presidential election, Plane, Microwave oven)

Fig. 2. Procedure of keyword acquisition using the user's Twitter

3. Words with high cosine similarity to words that the user is interested in.

To check the validity of the proposed approach, we examined two methods to observe what the user is currently interested in: "a method of acquiring and analyzing tweets from the user's Twitter data" and "a method of directly asking the user about what they are interested in." To ease the explanation, we labeled the first and second methods, TW and DA, respectively.

3.3 Acquisition of keywords that users are interested in

In this section, we explain how to prepare keywords that your users are interested in for topic selection. This research aims to provide topics that users are interested in but are unfamiliar with. To achieve this goal, the system searches for potential topics using keywords that users are interested in. We collected keywords that users are interested in. Two approaches were used to collect the data:

1. An approach to directly asking users their interests
2. An approach to analyzing what users' interests are from their Twitter data

The first data collection was done to evaluate the estimation ability of the proposed method. The second data collection was done to evaluate whether keywords could be collected automatically. The first way to ask users directly about their interests is the best way to select keywords from an accuracy standpoint. However, directly asking about the user's interests increases the burden on the user.

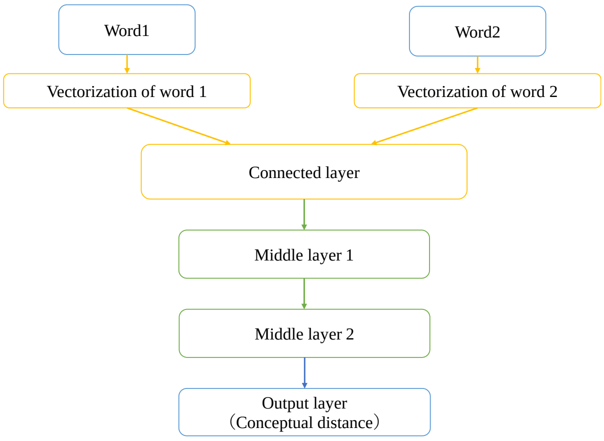

Fig. 3. Basic concept of the conceptual distance measurement model by machine learning

If keyword selection can be automated by the second method, the burden on users is expected to be reduced. The second method was investigated to assess the system automation potential.

For the first method, we asked users to choose 10 to 20 nouns as keywords of interest. The number of keywords was set so the subject could choose without burden. The keywords obtained from the subjects were 13–20.

Figure 2 shows the procedure for keyword preparation using the second method. In Fig. 2, we describe specific examples of tweets and obtain keywords. The examples are described in English and Japanese to show the details of the obtained tweets and keywords. In this procedure, the system obtains the latest 200 tweets from the user's tweets using the Twitter API to obtain the user's frequent tweets. We set the number of tweets to be acquired at 200 because it is the maximum number of tweets the Twitter API can acquire. The tweets obtained were used as data to generate users' keywords of interest. The system morphologically analyzed the acquired tweets and acquired nouns, which are candidate keywords. We used MeCab for morphological analysis [47]. Since keyword candidates included relatively new nouns, the Neologd dictionary [48] was utilized as the dictionary for MeCab. Neologd is a dictionary that also contains recently used words. To exclude inappropriate nouns as keywords, the system then removed everything except hiragana, katakana, and kanji. This action removed user IDs, alphabets, symbols, and numbers. We named the list of nouns resulting from these processes a keyword list for ease of explanation. The keyword list included both general nouns and proper nouns.

We assumed that the words users often tweet were likely to be interesting. Hence, we extracted the words the user was interested in from the keyword list. The system selected 25 words that appeared most frequently in the keyword list as keywords. The number of words for selection was experimentally determined by the results of previous experiments.

3.4 Tweet analysis

As a method of acquiring and analyzing tweets from Twitter data, we examined two methods, as follows:

1. A method for performing morphological analysis and extracting all nouns
2. A method of extracting topical words in tweets using IDF values.

Table 1. Fundamental physical constants

Model number	Intermediate layer 1	Intermediate layer 2
1	20	20
2	100	20
3	200	100
4	300	300

To ease the explanation, we labeled the first and second methods, the MA and IDF methods, respectively. In the MA method, the tweet obtained was subjected to morphological analysis using the morphological analyzer MeCab [47]. The Neologd dictionary was used for the dictionary [48]. General nouns and proper nouns, excluding user IDs, alphabets, symbols, numbers, links, RTs, and pictograms, were extracted from the obtained words.

In the IDF method, a topic word was extracted from each tweet using the topic word extractor. For the topic word extractor, we used the method of comparing the IDF values of each word following Mikami et al. [17]. The IDF value can be calculated as follows:

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}, \tag{1}$$

where D is the set of documents. $|\{d \in \mathrm{D}: t \in d\}|$ is the number of documents where the term t appears. In Mikami et al. [17], the word with the highest IDF value was adopted as the topic word. In addition, the words used as topical words are limited to general nouns and proper nouns.

Japanese Wikipedia data were used to calculate the IDF value. When calculating the IDF value, the Wikipedia article was treated as one document, and the IDF value of each word was calculated. During implementation, the IDF value in the tweet was calculated for each tweet, and the “general noun” and “proper noun” with the lowest IDF value were extracted as topic words.

3.5 Keyword search The system performed a keyword search to extract topical words related to the word of interest. As a search word, we set the words that the user answered or the words that frequently appeared among the words extracted from the tweets. The system acquired up to 100 tweets, the maximum number of tweets the Twitter API can acquire per word. Consequently, words were extracted from the obtained tweets according to the method in Section 3.4. After the keyword search, the system removed the words used in the keyword search from the extracted words. This is because the user is familiar with the words included in the tweet and the words answered by the user. The system evaluates the similarities of the remaining words. The details will be described in the next section.

3.6 Conceptual distance This section describes conceptual distance. Conceptual distance is defined as the number of concepts that go through to connect two words in a thesaurus. In this research, the Japanese WordNet [49] is used as a thesaurus dictionary. WordNet has a network structure, and the nodes are divided into “words” and “concepts.” Edges are divided into two categories: “sense,” which connects “words” and “concepts,” and “synlink,” which connects “concepts” and “concepts.” Since there is no link between “words,” there is always one or more “concepts” between “words.” If the number of “concepts” that pass through is small, then the words are conceptually close to each other, but if the number of “concepts” that pass through is large, the words are conceptually far from each other. The conceptual distance for the same word is zero. When there are two or more paths between words, the shortest distance is defined as the conceptual distance.

Our aim is to give users the topics that users are interested in but are not familiar with. To achieve this goal, we aim to measure the conceptual proximity of words and extract words that are close in simple terms but conceptually separated by a certain distance. There is no guarantee that things that are close are conceptually close. Hence, we employed conceptual distance. This is the reason why the distance has to be calculated by machine learning instead of using simple distance between word vectors.

3.7 Similarity evaluation The system evaluated the similarity between the five words used as keywords and the words obtained by keyword searches. Words were first vectorized using the trained model. Word2Vec was used to vectorize words [50–52]. In previous research, words with high cosine similarity were output, but conceptually similar words were output [20]. To solve the problem, we introduced the conceptual distance in this research, removed the words whose conceptual distance was too close to the keyword, and then output the ones with high cosine similarity.

3.8 Generation of conceptual distance measurement model by machine learning If two words to be measured are included in the Japanese WordNet, the conceptual distance between the words can be measured. However, if either word is not included in the dictionary, the conceptual distance cannot be measured. Additionally, research on topic extraction often deals with relatively new words that are not included in the dictionary.

To solve this problem, we created a conceptual distance generation model using machine learning to measure conceptual distances for all words, including words that are not included in the dictionary.

Figure 3 shows the outline of the model. The proposed model consists of a neural network; the input is two words, and the output is the conceptual distance (Fig. 3). The learning data are the conceptual distance data for all word combinations in Japanese WordNet. The model was trained using the vectorized version of both word pairs as the input and the normalized conceptual distance as the output. Two words are converted to 300-dimensional vectors using a trained model. The ReLU function was applied as the activation function of the intermediate layer, and the sigmoid function was applied as the activation function of the output layer.

When measuring the conceptual distance of an unknown word, we input the two target words into the trained model to obtain the conceptual distance. In the experiment, we prepared four models with different intermediate layers. We labeled them model1, model2, model3, and model4 to ease the explanation. The parameters of the models are shown in Table 1. To check the effect of the parameters, the comparison experiments were conducted [53]. Table 2 shows the results of 5 cross-validation by randomly dividing the training data into 5 parts. From the cross-validation results, it was confirmed that there was almost no difference in mean absolute error between the training data and the test data.

Table 2. Mean MAE values between true and estimated distance for training and test data.

Model	Middle layer 1	Middle layer 2	MAE in training data	MAE in test data
1	20	20	0.0225	0.0225
2	100	20	0.0202	0.0202
3	200	100	0.0191	0.0191
4	300	300	0.0187	0.0188

Table 3. Interests of Subject A

Financial engineering	Mahjong	Camp	Pot	Qualification
Ice	Movie	Shibuya	Aquarium	Tokyo Tower
Animation	Chidori	Oden	Donut	

Table 4. Interests of Subject B

DTM	APEX	FPS	Game	Manga
Karaage	Card	3D Printer	Computer	Tablet
Keyboard	Mouse	Smartphone		

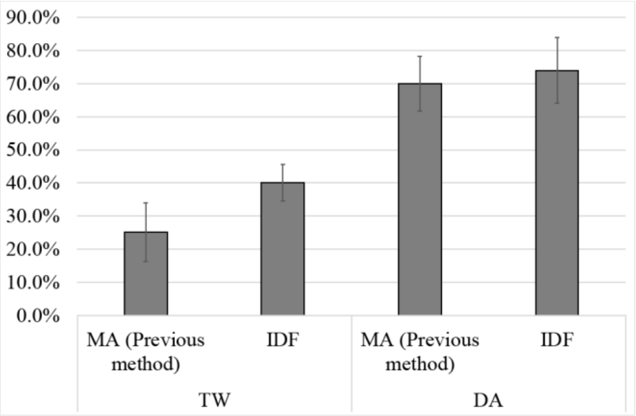

Fig. 4. Percentage of words that the user is interested in when the words with the highest similarity are provided as topical words. TW represents an approach to analyzing the user's interest from their Twitter data. MA represents a method for performing morphological analysis. IDF represents a method of extracting topical words from tweets using IDF values.

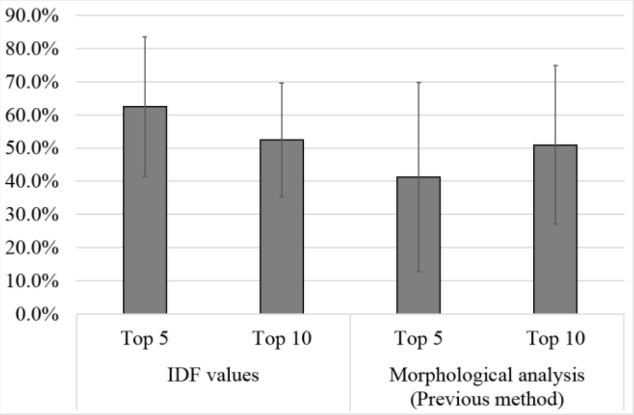

Fig. 5. Percentage of words that the user is interested in to the words obtained from tweets. Comparison of a method of performing morphological analysis and a method of extracting topical words in tweets using IDF values

Table 5. Percentage of users unfamiliar with the presented word with and without conceptual distance. Bold represents the case superior to the baseline DA represents an approach to directly ask users what they are interested in. TW represents an approach to analyzing what you are interested in from your Twitter data. MA represents a method for performing morphological analysis. IDF represents a method of extracting topical words in tweets using IDF values.

		DA		TW	
		MA	IDF	MA	IDF
Cosine similarity		45.5%	45.5%	63.1%	66.5%
WordNet	(4,7)	**50.4%**	**54.2%**	**69.7%**	61.0%
WordNet	(3,8)	**52.5%**	**48.3%**	**67.0%**	**68.8%**
WordNet	(2,9)	**52.9%**	**46.7%**	**68.4%**	66.3%
Model1	(4,7)	**62.0%**	**64.0%**	**65.4%**	63.5%
Model1	(3,8)	**55.5%**	**63.0%**	**67.4%**	**68.5%**
Model1	(2,9)	**47.0%**	**59.0%**	**68.6%**	**67.5%**
Model2	(4,7)	**64.5%**	**62.5%**	**69.4%**	63.5%
Model2	(3,8)	**59.5%**	**61.0%**	**69.3%**	66.0%
Model2	(2,9)	**54.5%**	**61.0%**	**65.5%**	**68.5%**
Model3	(4,7)	**60.5%**	**61.5%**	**65.9%**	62.0%
Model3	(3,8)	**55.5%**	**59.5%**	**63.8%**	61.0%
Model3	(2,9)	**54.5%**	**55.0%**	**64.4%**	65.0%
Model4	(4,7)	**58.5%**	**63.0%**	**64.5%**	65.5%
Model4	(3,8)	**53.5%**	**56.5%**	**65.4%**	63.0%
Model4	(2,9)	**50.5%**	**55.0%**	**64.5%**	64.0%

Table 6. Topic word examples. They are described in English and Japanese for easy understanding.

	Topic word examples
Baseline	Disney character
WordNet	Pompeii (ポンペイ) Dalmatian (101 匹わんちゃん) One piece (ワンピース)
Trained model	One piece (ワンピース) Spider-man (スパイダーマン) Witchy PreCure (魔法使いプリキュア！)

4. Experiments

4.1 Experimental setup An evaluation experiment was conducted to examine whether the proposed method could generate interesting but unfamiliar topics. Eight subjects participated in this experiment.

The experiments were conducted after approval from the University of Electro-Communications Ethical Committee, where the principal investigator belonged. First, we entered the subjects' tweets into the system and obtained the keyword word. Second, for each of the output words, the subjects were asked to evaluate "whether they were interested," "whether the content was familiar," and "whether the topic they were what they wanted to hear or talk about" on a four-point scale. The subjects were male and female university students in their 20s, the language of the tweets was Japanese, and the eight subjects' fields of interest were selected from their tweets. We showed two examples.

Table 3 and 4 show the examples of interests of subject A and subject B, respectively. We set questionnaires for each topic candidate. For example, financial engineering is a candidate for a subject topic. Here, we set the questionnaire as follows:

Regarding financial engineering,

Q1. Are you interested in this topic?

4. Very interested 3. A little interested 2. Not very interested 1. Not interested at all

Q2. Are you familiar with this topic?

4. Very familiar 3. A little familiar 2. Not very familiar 1. Not familiar at all

Q3. Do you want to listen to this topic?

4. I really want to listen to it 3. I want to listen to it a little. 2. I do not want to listen to it very much. 1. I do not want to listen to it at all.

Q4. Do you want to talk about this topic?

4. I really want to talk about it 3. I want to talk about it a little. 2. I do not want to talk about it very much. 1. I do not want to talk about it at all.

As comparison content, we compared two methods of collecting data: "subject's tweet" and "listen directly to the subject." Two methods of analysis for the obtained tweets were compared: "extract general/proper nouns by morphological analysis" and "extract topical words by the IDF method." The following 16 patterns were implemented for similarity evaluation:

1. Cosine similarity only (baseline in the past study [18])
2. Removed conceptual distances outside the range (4, 7), (3, 8), and (2, 9) using WordNet before using cosine similarity
3. Removed conceptual distances in the range (4, 7), (3, 8), and (2, 9) using the trained model (model1 to model4) before using cosine similarity

(*a*, *b*) shows the results of extracting topical words using cosine similarity after extracting words whose conceptual distance is *a* or more and *b* or less. A total of 64 comparisons were made.

We used cosine similarity to calculate the similarity between the two words. The cosine similarity between the word vector v_i and v_j is defined as follows:

$$\cos(v_i, v_j) = \frac{v_i \cdot v_j}{|v_i||v_j|} \qquad (2)$$

Words with high cos similarity are expected to be similar regarding the distributed hypothesis. After vectorizing words in Word2Vec, we calculated the cosine similarity between keywords and search results and selected highly similar words.

4.2 Evaluation by analysis method for obtained tweets An evaluation experiment was conducted to check the validity of the IDF method compared to the MA method (baseline in the past study [20]). To check the validity, the average values of the MA and IDF methods for 64 combinations were calculated for the keywords acquired by the DA and TW methods, respectively. Fig.4 shows the percentage of words the user is interested in when the words with the highest similarity are provided as topical words with standard deviation. In both cases, the results of the IDF method were superior to those of the MA method. Fig. 5 shows the percentage of words extracted from the subject's tweets where they answered that they were interested when using their tweets. It was confirmed that the subjects were more interested in the words acquired using the IDF method.

4.3 Evaluation of the effect of conceptual distance We conducted a comparison experiment to evaluate the effect of conceptual distance. Cosine similarity was used in our previous method as the baseline [20]. We aimed to select topics that users were interested in but unfamiliar with. Cosine similarity is used in [20] to measure word similarity. Previous research was relatively good from the user's viewpoint, but the accuracy was bad from the viewpoint of "not familiar with it". The conceptual distance was introduced in this study to provide unfamiliar topics. Objects with close conceptual distances are semantically similar, so users are expected to be familiar with them. The proposed method presents unfamiliar words to the user by removing words with close conceptual distances. Table 5 shows the percentage of users unfamiliar with the presented word when the DA and TW methods were used. It shows all 64 cases, that is, removed words with conceptual distances outside the range (4, 7), (3, 8), and (2, 9) using WordNet or the trained model before using cosine similarity. In Table 5, cosine similarity shows the baseline results of extracting topical words using only cosine similarity. Bold in Table 5 shows the results superior to the baseline.

4.4 Concrete examples We showed some concrete examples to clarify the kinds of words suggested. Table 6 shows an example of the output word for a certain subject. Here, the keyword that the user was interested in was "Disney". In Table 6, the baseline shows the topic word candidate when only cosine similarity was used. WordNet showed the topic word candidates when WordNet was used before outputting those with high cosine similarity. The trained model showed the topic word candidates when the trained model was used instead of WordNet.

As shown in Table 6, a conceptually similar word, Disney character, was proposed for Disney when using only cosine similarity. Conversely, when we used WordNet and the trained model, we obtained various topical words that are conceptually different from Disney. Moreover, we obtained words like Witchy PreCure that were not included in WordNet using the trained model.

5. Conclusion

In this paper, we investigated a method for topic selection to provide topics that users were interested in but unfamiliar with. The following improvements have been made to our previous studies.

We proposed a topic word extraction method using IDF to provide highly interesting topics to users compared to past research [20]. Furthermore, we introduced conceptual distances to help users provide fewer familiar topics. We proposed a conceptual distance measurement model using machine learning so that words not included in the dictionary could be handled. The experiment confirmed that the user's topic of interest could be extracted using the IDF method. Moreover, a topic unknown to the user can be presented using conceptual distance. Generally, the improved system can provide users with topics they are interested in but are unfamiliar with.

Since the proposed method is for acquiring topic words that the user is interested in but is unfamiliar with, it is necessary to separately consider how to output sentences from the topic words.

A simple method is to discuss whether the user is interested in the output topic word. In past research, there was a method of obtaining an utterance candidate sentence by inputting a topic word and talking to the user using the utterance sentence [54]. In the future, whether these methods can be applied to topical word output in this study must be examined. Advertising selection is also an application candidate for our method.

Subsequently, we will consider applying this method to a topic word determination system that uses dialogue data with users.

We uploaded our source codes to GitHub and provided its URL as the reference to increase the credibility of the study [55].

References

(1) L. Velikovich, I. Williams, J. Scheiner, P. Aleksic, P. Moreno, and M. Riley : Semantic Lattice Processing in Contextual Automatic Speech Recognition for Google Assistant, Interspeech2018, pp.2222-2226 (2018).

(2) A. Michaely, C. Parada, F. Zhang, G. Simko, and P. Aleksic : Keyword Spotting for Google Assistant Using Contextual Speech Recognition, IEEE Automatic Speech Recognition and Understanding Workshop (2017).

(3) M. Assefi, G. Liu, M. P. Wittie, and C. Izurieta : "An experimental evaluation of Apple Siri and Google speech recognition", Proc. of the 2015 ISCA SEDE.

(4) H. Chen, et al. : "A survey on dialogue systems: recent advances and new frontiers", in ACM SIGKDD Explorations Newsletter, Vol.19, No.2, pp.25-35, (2017).

(5) J. Nii, T. Young, V. Pandelea, F. Xue, and E. Cambria : "Recent advances in deep learning based dialogue systems: a systematic survey," Artificial Intelligence Review, 2022.

(6) R. Nishimura, et al. : "Web-based environment for user generation of spoken dialog for virtual assistants", J Audio Speech Music Proc. 2018, 17 (2018).

(7) C. Montenegro, et al. : "A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly", Multimodal Technol. Interact. Vol.3, pp.52 (2019).

(8) M. Henderson, et al. : "The Second Dialog State Tracking Challenge. In SIGDIAL, pp.263-272 (2014).

(9) P. Budzianowski, et al. : "MultiWOZ - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling", in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp.5016-5026, (2018).

(10) B. Liu et al. : "Content-oriented user modeling for personalized response ranking in chatbots", in IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.26, No.1, pp.122-133 (Jan. 2018).

(11) F. Patel, R. Thakore, I. Nandwani, and S. K. Bharti : "Combating depression in students using an intelligent ChatBot: a cognitive behavioral therapy," 2019 IEEE 16th India Council International Conference (INDICON), Rajkot, India, 2019, pp. 1-4.

(12) S. Srivastava and T. V. Prabhakar, "Desirable Features of a Chatbot-building Platform", 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI), Irvine, CA, USA, 2020, pp. 61-64.

(13) E. H. Wu, C. Lin, Y. Ou, C. Liu, W. Wang, and C. Chao : "Advantages and Constraints of a Hybrid Model K-12 E-Learning Assistant Chatbot," in IEEE Access, Vol.8, pp.77788-77801 (2020).

(14) T. Kubota et al. : "Implementation and evaluation of chat-oriented dialog system for an android robot in live streaming media in which users can speak at any time", J. of the Japanese Society for Artificial Intelligence, Vol.33, No.1, pp.1-13 (2018).

(15) Z. Miyashita et al. : "A robot in a shopping mall that affectively guide customers", Journal of the Robotics Society of Japan, Vol.26, No.7, pp.821-832, (2008).

(16) M. A.-Chenaghlu, M. R. F. Derakhshi, L. Farzinvash, M. A Balafar, and C. Motamed : "Topic Detection and Tracking Techniques on Twitter: A Systematic Review", Complexity, vol.2021, Article ID 8833084, pp.15, (2021).

(17) Y. Mikami et al. : "Topic expansion method considering randomness for dialogue system", J. of Japan Society of Kansei Engineering, Vol.17, No.3, pp.365-373 (2018).

(18) N. Kondo and O. Uchida : "LDA based interest estimation method using Twitter", Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing, NLP2015-P3-32 (2015).

(19) J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi : "Short and tweet: experiments on recommending content from information streams," CHI '10, pp.1185-1194 (2010).

(20) Y. Sakai and M. Matsumoto : "Twitter-based selection of topics that users are interested in but are not familiar with", 2021 IEEE/SICE International Symposium on System Integration, pp.769-774 (2021).

(21) Y. Sakai and M. Matsumoto : "On Selection of Topics That Users are Interested in but are Not Familiar with", 2022 International Power Electronics Conference (IPEC2022), pp.911-915 (2022).

(22) J. Deriu, A. Rodrigo, A. Otegi, et al. "Survey on evaluation methods for dialogue systems", Artif Intell Rev Vol.54, pp.755–810 (2021).

(23) S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman : "Indexing by latent semantic analysis", J Am Soc Inf Sci Vol.41, No.6, pp.391 (1990).

(24) T. Hofmann : Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp.50-57 (1999).

(25) D. Lee and H. Seung : Algorithms for non-negative matrix factorization. In: Proceedings of the advances in neural information processing systems 13 (NIPS 2000). Denver, CO, USA, pp.556-562 (2000).

(26) D. Blei, A. Ng, and M. Jordan : "Latent dirichlet allocation", J Mach Learn Res Vol.3, pp.993-1022 (2003).

(27) A. McCallum, A. Corrada-Emmanuel, and X. Wang : "The author–recipient–topic model for topic and role discovery in social networks: experiments with enron and academic email, Workshop on Link Analysis, Counterterrorism and Security, pp.33-44 (2005).

(28) A. McCallum, X. Wang, and A. Corrada-Emmanuel : "Topic and role discovery in social networks with experiments on enron and academic email", J Artif Intell Res Vol.30, pp.249-272 (2007).

(29) X. Wang and A. McCallum, Topics over time: A non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, KDD '06, pp.424-433 (2006).

(30) D. M. Blei and J. D. Lafferty : "A correlated topic model of science", Ann Appl Stat Vol.1, pp.17-35 (2007).

(31) L. Dietz, S. Bickel, and T. Scheffer : Unsupervised prediction of citation

influences. In: Proceedings of the 24th international conference on machine learning. ACM, New York, NY, USA, ICML '07, pp.233-240 (2007).

(32) Y. Liu, A. Niculescu-Mizil, and W. Gryc : Topic-link LDA: Joint models of topic and author community. In: Proceedings of the 26th annual international conference on machine learning. ACM, New York, NY, USA, ICML '09, pp.665-672 (2009).

(33) D. Ramage, D. Hall, R. Nallapati, and C. D. Manning : Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing: volume 1–volume 1, association for computational linguistics. Stroudsburg, PA, USA, EMNLP '09, pp.248-256 (2009).

(34) K. W. Prier, M. S. Smith, C. Giraud-Carrier, and C. L. Hanson : Identifying health-related topics on Twitter. In: Salerno J, Yang SJ, Nau D, Chai SK (eds) Proceedings of the social computing, behavioral-cultural modeling and prediction: 4th international conference, SBP 2011, pp.18-25 (2011).

(35) K. Kireyev, L. Palen, and K. Anderson, Applications of topics models to analysis of disaster-related Twitter data. In: Proceedings of the NIPS workshop on applications for topic models: text and beyond. Whistler, Canada, Vol.1 (2009).

(36) C. Zhang, S. Lu, C. Zhang, X. Xiao, Q. Wang, and G. Chen : "A novel hot topic detection framework with integration of image and short text information from twitter", IEEE Access Vol.7, pp.9225-9231 (2019).

(37) X. H. Phan, L. M. Nguyen, and S. Horiguchi : Learning to classify short and sparse text and web with hidden topics from large-scale data collections. In: Proceedings of the 17th international conference on WorldWideWeb, pp.91-100, (2008).

(38) J. Weng, E. P. Lim, J. Jiang, and Q. He, Twitterrank: finding topic-sensitive influential Twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. ACM, pp.261-270 (2010).

(39) X. H. Phan, C. T. Nguyen, L. M. Le, D. T. Nguyen, S. Horiguchi, and Q. T. Ha : "A hidden topic-based framework toward building applications with short web documents", IEEE Trans Knowl Data Eng Vol.23, No.7, pp.961-976 (2011).

(40) X. Hu, N. Sun, C. Zhang, and T. S. Chua, Exploiting internal and external semantics for the clustering of short texts using world knowledge. In: Proceedings of the 18th ACM conference on information and knowledge management. ACM, New York, NY, USA, CIKM '09, pp.919-928 (2009).

(41) J. Deriu, A. Rodrigo, A. Otegi, et al., "Survey on evaluation methods for dialogue systems", Artif Intell Rev Vol.54, pp.755-810 (2021).

(42) A. Lavie and M. J. Denkowski : "The meteor metric for automatic evaluation of machine translation", MachTransl Vol.23, No.2-3, pp.105-115 (20090.

(43) F. Charras, G. Dubuisson Duplessis, V. Letard, A. L. Ligozat, and S. Rosset, Comparing system-response retrieval models for open-domain and casual conversational agent. In: Workshop on Chatbots and Conversational Agent Technologies (2016).

(44) D. G. Dubuisson, V. Letard, A. L. Ligozat, and S. Rosset : Purely corpus-based automatic conversation authoring. In: Proceedings of the tenth international conference on language resources and evaluation, European Language Resources Association (ELRA), Paris, France, LREC (2016).

(45) X. Yan, J. Guo, S. Liu, X. Cheng, and Y. Wang : Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In: Proceedings of the SIAM international conference on data mining (SIAM 2013) (2013).

(46) R. Lowe, N. Pow, I. V. Serban, L. Charlin, C. W. Liu, and J. Pineau : "Training end-to-end dialogue systems with the ubuntu dialogue corpus", Dialogue Discourse, Vol.8, No.1, pp.31-65 (2017).

(47) https://taku910.github.io/mecab/

(48) https://github.com/neologd/mecab-ipadic-neologd

(49) http://compling.hss.ntu.edu.sg/wnja/index.en.html

(50) T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations", Proc. of NAACL-HLT, pp.746-751 (2013).

(51) T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, Proceedings of the International Conference on Learning Representations (2013).

(52) T. Mikolov, I. Sutskever, K. Chen, G S Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, Proceedings of the 26th International Conference on Neural Information Processing Systems, pp.3111-3119 (2013).

(53) Y. Sakai, M. Matsumoto, "Estimating the conceptual distance between unknown words using machine learning," Proc. of 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems, (2022).

(54) M. Inaba et al., "Candidate utterance acquisition method for non-task-oriented dialogue system from Twitter," J. of the Japanese Society for Artificial Intelligence, Vol.29, No.1, pp.21-31 (2014).

(55) https://github.com/MrSakaikun/TopicSelectSystem

Yuya Sakai (Non-member) received his B.E. and M.E. from the University of Electro-Communications, Japan, in 2019 and 2021, respectively. His research interests include dialog systems and machine learning.

Mitsuharu Matsumoto (Non-member) is currently an associate professor with the University of Electro-Communications, Japan. He received his B.E. in Applied Physics and M.E. and Dr. Eng. in Pure and Applied Physics from Waseda University, Tokyo, Japan, in 2001, 2003, and 2006, respectively.

His research interests include acoustical signal processing, image processing, pattern recognition, self-assembly, human–robot interaction, and robotics. He received the Ericsson Young Scientist Award from Nippon Ericsson K.K, Japan, and the FOST Kumada Award, in 2009 and 2011, respectively. He has published over a hundred of journal and international conference papers. He is a member of the Institute of Electrical and Electronic Engineers (IEEE), the Institute of Electronic, Information and Communication Engineers (IEICE), and the Society of Instrument and Control Engineers (SICE).