ISSN 2007-9737

Assigning Character Strings to Links for Construction of Concept Networks

Masaki Murata, Yuhei Kubo

Tottori University, Faculty of Engineering,
Japan

murata@tottori-u.ac.jp, s102014@ike.tottori-u.ac.jp

Abstract. In this paper, we propose a method of extracting character strings from newspaper articles and adding them to the links of a word network to express node relationships. Attaching such strings to links enables us show the word relationships. When we evaluated the results in terms of the mean reciprocal rank (MRR), Top-1 accuracy rate, and Top-5 accuracy rate and considered answers with additional or missing information to be correct, the MRR was about 0.7, the Top-1 accuracy rate was about 0.6, and the Top-5 accuracy rate was about 0.9.

Keywords. Concept network, link, character string, MRR.

1 Introduction

In recent years, vast and increasing numbers of electronic texts have been posted on the Internet, so we need to find ways of automatically extracting useful information from them. Doen et al. [4] proposed a method of extracting relationship information items based on identifying specific keywords in the text and creating a network based on them, then constructed such a network based on the keyword “earthquake.” They found that the network included nodes for unrelated things and proposed a method of automatically deleting them.

Although they constructed a network of concepts, their aim was not to construct a dictionary, such as WordNet (word knowledge), but instead to construct a network that could be used to generate ideas and understand concepts. For example, their constructed concept network about earthquake includes a relation between an earthquake and a nuclear power plant.

The construction of concept networks is also useful for summarization of many documents [2, 10, 11]. The method can make a network that is useful for grasping information related to a concept from a lot of documents related to the concept.

However, their network did not include any node relationship information, making it difficult to understand the relationships. In this study, we therefore propose a method of extracting character strings from newspaper articles and assigning them to links in word networks to express node relationships. By adding character strings to links to indicate the relevant relationships, we can obtain more detailed information from the resulting word network. The aim of this study is to make word networks more useful by making the node relationships easier to understand. This study was conducted in Japanese.

There are many studies related to relation extraction [1, 3, 5, 6, 8, 12, 13, 14]. These studies extract categories of relation between words and words with a certain relation, where the relation includes part-of, entity-place, person-company and so on. In contrast, our study extracts expressions that explain the relationship between word pairs in some detail.

The main contributions of our study are as follows:

— In order to deal with the issue that word networks do not include node relationship information and hence that these relationships are difficult to understand, we attach character strings indicating the node relationships to the links in the word network.

doi: 10.13053/CyS-26-1-4145

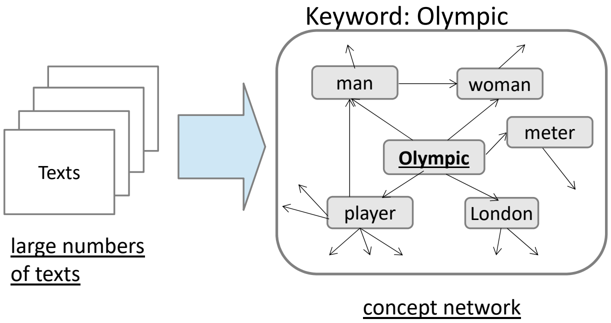

Fig. 1. Network construction overview (Source: Figure 1 of the paper [4])

- Attaching character strings to the links makes the relationships between words easier to understand.
- We evaluate the output of the proposed method using the mean reciprocal rank (MRR), Top-1 accuracy rate, and Top-5 accuracy rate. Here, we found that they were about 0.7, 0.6, and 0.9, respectively, based on considering character strings with unnecessary or missing information to be correct.
- We consider two main methods of extracting character strings. The first uses commas and periods (i.e., punctuation marks dividing clauses) as delimiters, while the other uses periods only (punctuation marks dividing sentences). After comparing the experimental results for both methods, we found mixed results: the method using both commas and periods produced better results when only strings that correctly expressed the word relationships were considered to be correct, but the method using periods only performed better if we also allowed strings with additional information. In addition, if we allowed strings with both additional and missing information, both methods performed very similarly.

2 Network Construction

2.1 Overview

Here, we construct word networks using Doen et al.'s method [4], based on a large newspaper article dataset (Figure 1). The procedure is as follows:

1. Select a keyword[1] that expresses the main concept around which the network is to be constructed.
2. Extract all articles from the newspaper dataset that include the keyword.
3. Apply morphological analysis to the resulting articles and extract words related to the keyword.
4. Create text nodes for the five words that are most closely related to the keyword and connect them to the keyword node. Here, we use a related word extraction method based on term frequency–inverse document frequency (TF-IDF) to identify these five words.[2]
5. Select each new word added to the network in turn as the new keyword and repeat from Step 2 to expand the network.

In this work, we also remove unrelated words from the network by adding additional procedures to Steps 4 and 5 by referring to the method proposed by Doen et al. The method is explained in detail in the following sections.

[1] The first keyword corresponds to the top node of a network. The first keyword can be selected by a user freely. A user should select a word that corresponds to the contents of the network he wants to create as the first keyword.

[2] In the study, we used TF-IDF to gather related keywords. However, we can also use word embeddings such as word2vec for gathering related keywords. In the future, we would like to expand our method so as to use word2vec.

2.2 Extracting Candidate Nodes

Denote the initial keyword by k. First, we extract the articles that include k, and denote the resulting article set as S. Next, we use morphological analysis (specifically, ChaSen [7]) to extract the nouns from S. Here, we eliminate single-character words and those consisting only of hiragana characters (Japanese functional characters) or figures, because such words are unlikely to be important. Then, we use the remaining words as candidate nodes.

2.3 Selecting Nodes using TF-IDF-based Related Word Extraction

From the candidate nodes, we select the ones to actually add to the network via TF-IDF-based related word extraction. Specifically, we score the words using a TF-IDF based method and identify the five highest-scoring ones as being most closely related to the keyword. Then, we add the highest-scoring candidate words to the network as nodes, and use the TF-IDF scores as edge weights.

The TF-IDF based related word extraction method produces scores indicating the importance of particular words (candidate nodes) in the extracted articles, which are calculated as follows:

$$TF-IDF = tf_t \times \log \frac{N}{df_t}. \tag{1}$$

Here, tf_t is the frequency with which candidate node word t appears in the extracted articles, while df_t is the number of articles where the candidate node word t appears in the entire dataset and N is the total number of articles in the dataset.

Equation 1 uses df_t to give low weights to less-important words that appear in many articles, and conversely gives high weights to significant words that only appear in a small number of articles.

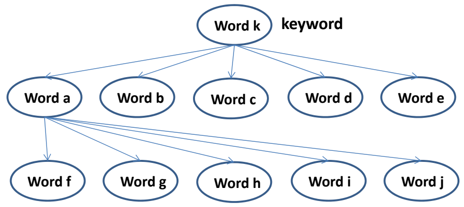

Fig. 2. Example of building a network by extracting nodes (Source: Figure 3 of the paper [4])

2.4 Expanding the Network

First, we extract five words based on the keyword k, as described in Section 2.3, and connect them to k. Next, we expand the network by extracting five additional words, based on using the first word a of the initial five as a new keyword, and connecting them to a (Figure 2). Then, we continue to expand the network by iteratively repeating this process.

2.5 Deleting Unrelated Words

Next, we use the topic-restricted extraction method proposed by Doen et al. to delete unrelated words from the network. This involves one change to the method given in Section 2.1. Specifically, while extracting articles by repeating Step 5, we only extract those that include both the initial and current keywords. Since this means we are focusing on articles that include the initial keyword, we are likely to obtain words related to it, and unlikely to extract unrelated words (Figure 3).

3 Proposed Method: Selecting Character Strings to Attach to Links

In order to make the node relationships in the word network easier to understand, we attach a character string indicating the relationship between the corresponding words to each link. For example, we might take the newspaper article dataset and the input word pair “universe” and “exploration” (Figure 6 in Section 4.1), and output the character string to attach to the link between the words. Figure 4 shows an example of assigning a string to a link in a word network. We select the character string as follows:

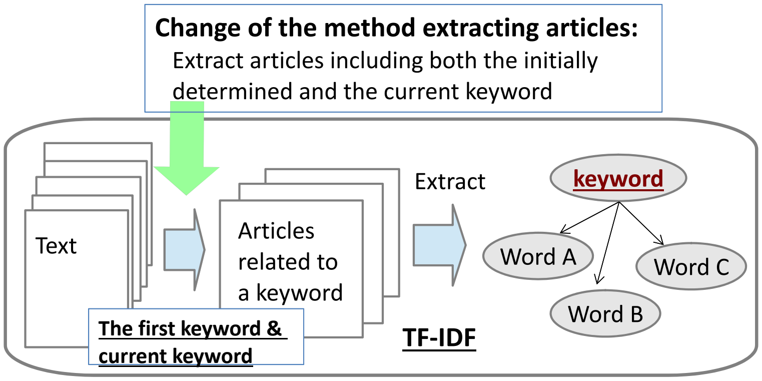
Change of the method extracting articles:
Extract articles including both the initially determined and the current keyword
Text
Articles related to a keyword
The first keyword & current keyword
Extract
TF-IDF
keyword
Word A
Word B
Word C

Fig. 3. Overview of topic-restricted extraction (Source: Figure 4 of the paper [4])

Table 1. Examples of extracting character strings A and B

Word pair	Character string A	Character string B	Original character string
girisha (Greece), *kokusai* (goverment bonds)	*no* ('s)	*chugoku wa zaisei saiken ni torikumu girisha no kokusai wo kounyu shi* (China purchases Greece's government bonds to assist fiscal rebuilding)	*chugoku wa zaisei saiken ni torikumu girisha no kokusai wo kounyu shi, yuuro bouei ni kyouryoku suru shisei wo shimesu nado oushuu eno eikyou ryoku wo kakudai shiteiru* (China purchases Greek government bonds to assist fiscal rebuilding, and indicates a desire to cooperate over the euro, including expanding influence in Europe.)
toyota (Toyota), *suiso* (hydorogen)	*jidousha wa* (automobiles)	*toyota jidousha wa suiso de ugoku nenryou denchi sha wo 2014 nendo ni kokunai de hatsubai to happyou* (Toyota announces the sale of hydrogen-powered fuel cell vehicles in Japan in 2014.)	*toyota jidousha wa suiso de ugoku nenryou denchi sha wo 2014 nendo ni kokunai de hatsubai to happyou. shihan wa seika hatsu to naru mitooshi* (Toyota automobiles announces the sale of hydrogen-powered fuel cell vehicles in Japan in 2014. This market is expected to be the world's first)

Table 2. Evaluation criterion and example for "Good"

Criterion	The output appropriately indicates the relationship between the two words.
Example	*wakata koichi uchu hikoushi: ISS sencho* (Koichi Wakata astronaut (universe flying pilot): Captain of the ISS)

Table 3. Evaluation criterion and example for "OK1"

Criterion	The output appropriately indicates the relationship between the two words, but includes additional information.
Example	*nihonjin hatsu no sencho wo tsutometa wakata kouichi uchu hikoushi (50) wa 14 nichi gozen 7 ji 58 hun* (Wakata Koichi astronaut (universe flying pilot) (50), the first Japanese captain, at 7:58 am on the 14th)

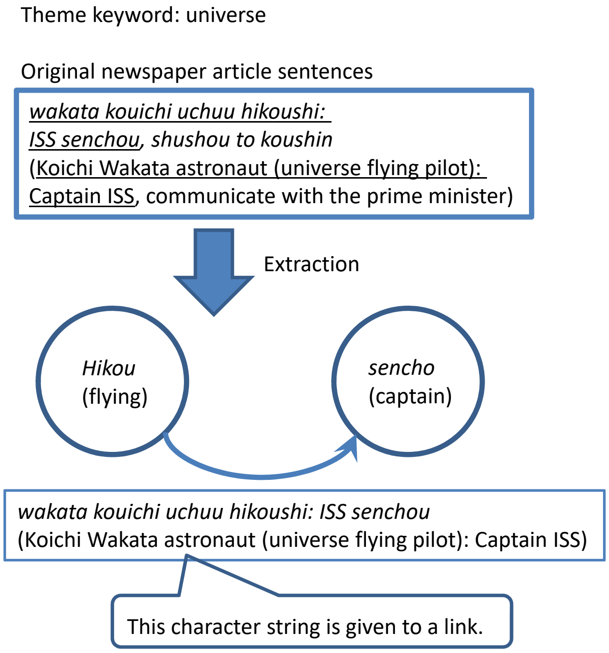
Theme keyword: universe
Original newspaper article sentences
wakata kouichi uchuu hikoushi:
ISS senchou, shushou to koushin
(Koichi Wakata astronaut (universe flying pilot):
Captain ISS, communicate with the prime minister)
Extraction
Hikou
(flying)
sencho
(captain)
wakata kouichi uchuu hikoushi: ISS senchou
(Koichi Wakata astronaut (universe flying pilot): Captain ISS)
This character string is given to a link.

Fig. 4. Example of assigning a string to a word network link

Table 4. Evaluation criterion and example for "OK2"

Criterion	The output appropriately indicates the relationship between the two words, but it is lacking information needed to make the relationship easier to understand.
Example	*nihonjin uchu hikoushi no sencho ga tanjo shiteiru* (The Japanese astronaut (universe flying pilot) captain appears)

Table 5. Evaluation criterion and example for "Bad"

Criterion	The output does not appropriately indicate the relationship between the two words.
Example	*kounin no sencho to natta bei koukuu uchu kyoku (NASA) no sutibun swanson hikoushi wa kouichi no riidaashippu wa subarashikatta to tatae* (Steven Swanson astronaut (universe flying pilot) of the National Aeronautics and Space Administration (NASA), who was the following captain, said "Koichi's leadership was wonderful")

1. From the dataset, extract a character string (character string A) from between the two words in a particular document.

2. Extract a character string (character string B) that includes string A and is delimited by periods and commas. Table 1 gives two examples of extracting character strings A and B.

3. Extract the highest-priority character string (character string C) from string B. High-priority strings are defined as those that either occur frequently or are short. This procedure (extracting string C) is repeated for all possible character strings A.

4. Select the highest-priority character string from among the strings C extracted in Step 3.

5. Attach the selected character string to the link.

The character string priorities are determined by one of the following three equations. Equation 2 prioritizes frequently-occurring strings, while Equation 3 prioritizes short strings and Equation 4 focuses on the ratio between frequency and length:

$$Priority = (Frequency \times 10000) - Length, \quad (2)$$

$$Priority = -(Length \times 10000) + Frequency, \quad (3)$$

$$Priority = \frac{Frequency}{Length}. \quad (4)$$

4 Experiments

4.1 Methods

In these experiments, we constructed word networks for the theme keywords “Toyota,” “universe,” and “Greece.” These networks consisted of 191, 228, and 99 word pairs, respectively. To build the networks for “Toyota” and “universe,” we used 102,547 articles taken from the Mainichi Newspaper (all from 2014). To build the network for “Greece,” we used 92,807 articles taken from the Mainichi Shimbun (all from 2010). Figures 5-7 give the networks for “Toyota,” “universe,” and “Greece,” respectively. These show that all the networks used here consisted of four levels.

4.2 Evaluation based on Human Judgment

Next, we evaluated whether or not the character strings given to the network links were appropriate. For this, we used 20 randomly chosen word pairs from each network (“Toyota,” “universe,” and “Greece”), for a total of 60 word pairs. In addition, ten randomly chosen newspaper articles including each word pair were used as reference data.

The pairs were then evaluated by a human participant against the top five highest-priority strings (as determined in Section 3), as determined by each of the three priority equations. The participant evaluated them on a four-step scale after consulting the reference newspaper articles. Hereafter, we will describe the methods embodied by Equations 2, 3, and 4 as “high-frequency,” “short,” and “ratio,” respectively. Tables 2–5 show the evaluation criteria and examples representing each of the four possible grades.

The example in Table 2 was evaluated as “Good” because the character string essentially says “astronaut Wakata Koichi has become the ISS captain,” which was judged to appropriately indicate the relationship between “flying” and “captain.” Similarly, the example in Table 3 was evaluated as “OK1” because, although the character string appropriately indicates the relationship between the two words, it also includes the additional phrase “at 7:58 am on the 14th.”

Next, the example in Table 4 was evaluated as “OK2” because, while the character string does indicate the relationship between the two words, it is missing information such as the person’s name and thus does not make the relationship as easy to understand as possible. Finally, the example in Table 5 was evaluated as “Bad” because the character string differed significantly from the correct information, as determined from the reference data, so it was judged as not appropriately indicating the relationship.

4.3 Evaluation Using the Mean Reciprocal Rank (MRR)

After conducting the manual evaluation described in Section 4.2, we evaluated the results using the MRR, based on the highest-ranked correct answer among the top five highest-priority results for each target. We calculated the MRR as follows:

$$MRR = \frac{\sum_{i=1}^{N} 1/r_i}{N}, \quad (5)$$

where N is the total number of targets to be evaluated and r_i is the rank of the highest-ranked correct answer produced for target i. In this study, since we output the top five results, $1 \leq r_i \leq 5$. Here, we considered three evaluation criteria: only “Good” answers are correct; “Good” and “OK1” answers are correct; and “Good,” “OK1,” and “OK2” answers are correct.

4.4 Evaluation Using the Top-n Accuracy Rate

Next, we evaluated the results based on the Top-n accuracy rates by assessing which of the top five highest-priority outputs corresponds to the correct answer. In the Top-n accuracy rate, a value of 1 is

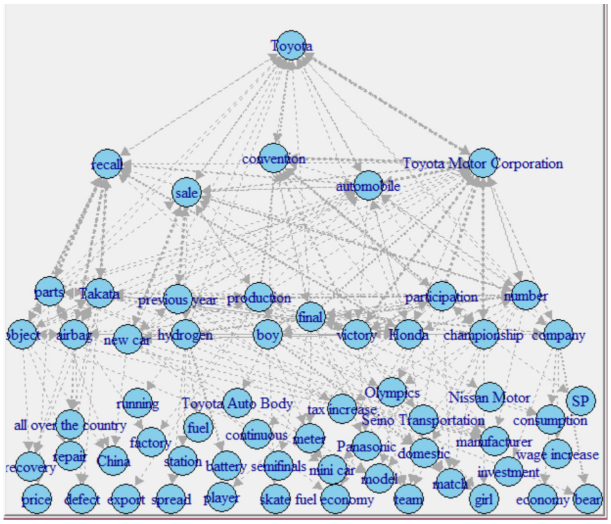

Fig. 5. Word network for "Toyota"

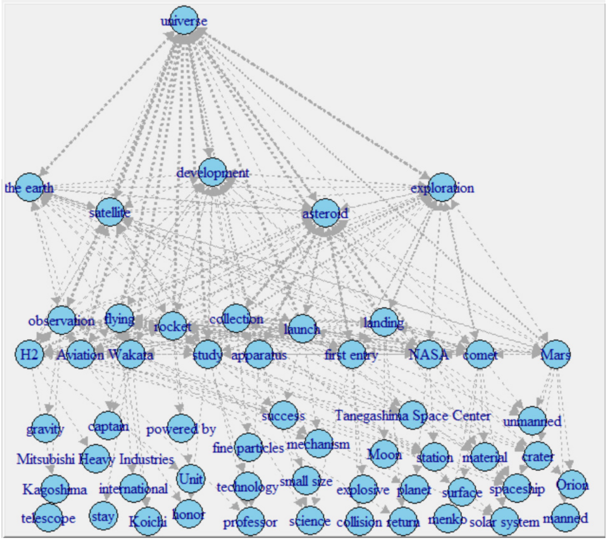

Fig. 6. Word network for "universe"

given in each question when the top n candidate answers contain the correct answer.

The Top-n accuracy rate is the result dividing the sum of the given values by the number of questions. In this study, we evaluated the Top-1 and Top-5 accuracy rates.

As for the MRR evaluation, we considered three correctness criteria: only "Good" answers are correct; "Good" and "OK1" answers are correct; and "Good," "OK1," and "OK2" answers are correct.

4.5 Evaluation of the First String Extraction Method (Periods and Commas)

In this section, we evaluate the string extraction method based on using periods and commas as delimiters. First, we randomly selected 20

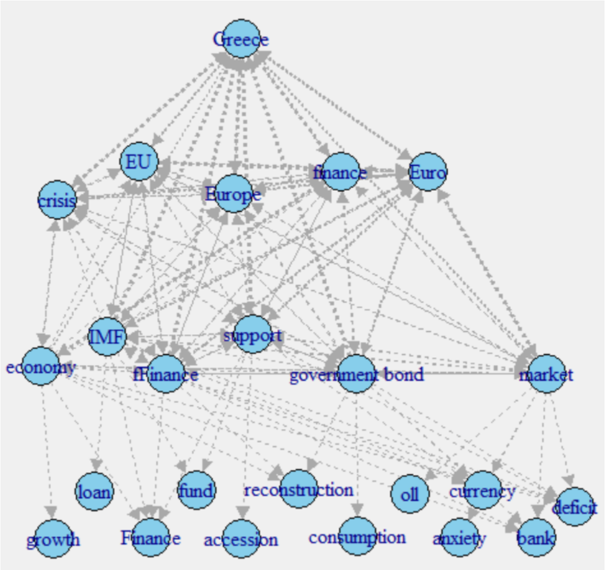

Fig. 7. Word network for “Greece”

Table 6. MRR evaluation results (periods and commas)

	Good	Good /OK1	Good/ OK1/OK2
High-frequency	0.30	0.44	0.68
Short	0.21	0.31	0.53
Ratio	0.28	0.41	0.68

Table 7. Top-1 accuracy rates (periods and commas)

	Good	Good /OK1	Good/ OK1/OK2
High-frequency	0.17	0.32	0.55
Short	0.12	0.15	0.35
Ratio	0.18	0.30	0.55

Table 8. Top-5 accuracy rates (periods and commas)

	Good	Good /OK1	Good/ OK1/OK2
High-frequency	0.53	0.68	0.90
Short	0.37	0.57	0.83
Ratio	0.47	0.63	0.88

word pairs each from the “Toyota,” “universe,” and “Greece” networks, for a total of 60 word pairs. For each word pair, we used the proposed method to extract the five highest-priority strings, then evaluated them in terms of their MRR, Top-1 accuracy rate, and Top-5 accuracy rate.

The experiments were carried out as described in Section 4.1, while the evaluations were carried out as described in Sections 4.2, 4.3, and 4.4. Table 6 shows the MRR evaluation results, while Table 7 shows the Top-1 accuracy rates and Table 8 shows the Top-5 accuracy rates. In addition, Tables 9–11 show some example outputs. “High-frequency,” “Short,” and “Ratio” corresponds to the use of Equations 2 to 4.

4.6 Evaluation of the Second String Extraction Method (Periods Only)

In this section, we evaluate the string extraction method based on using just periods as delimiters. The experimental conditions were as before (Section 4.1), and the evaluations were carried out as described in Sections 4.2, 4.3, and 4.4. Here, we used the same 60 word pairs for evaluation as in Section 4.5. Table 12 shows the MRR evaluation results, while Table 13 shows the Top-1 accuracy results, while Table 13 shows the Top-1 accuracy rates and Table 14 shows the Top-5 accuracy rates. Tables 15–17 show some example outputs.

Table 9. Example outputs for the “Toyota” network and the word pair “Takata” and “repair,” using commas and periods as delimiters

Priority	Output	Evaluation
High-frequency	*jidousha buhin ohte takata sei ea baggu no rikouru (kaishu/mushoushuri) mondai de* (in a problem leading to the recall (collection / free repair) of airbags made by Takata, a major auto parts company)	Good
Short	*takata sei ea baggu: 474 man dai shuuri wo* (Airbags by Takata: 4.74 million units repaired)	OK2
Ratio	*jidousha buhin ohte takata sei ea baggu no rikouru (kaishu/mushoushuri) mondai de* (in a problem leading to the recall (collection / free repair) of airbags made by Takata, a major auto parts company)	Good

Table 10. Example outputs for the “universe” network and the word pair “rocket” and “asteroid,” using commas and periods as delimiters

Priority	Output	Evaluation
High-frequency	*shou wakusei tansaki “hayabusa” wo noseta shuryoku roketto H2A26 gouki wo uchiageta.* (They launched an H2A26 main rocket with the “Hayabusa 2” asteroid explorer.)	Good
Short	*H2A roketto de uchiagerareru shou wakusei tansaki “hayabusa 2”* (Asteroid explorer ”Hayabusa 2” launched using an H2A rocket)	Good
Ratio	*shou wakusei tansaki “hayabusa” wo noseta shuryoku roketto H2A26 gouki wo uchiageta.* (They launched an H2A26 main rocket with the “Hayabusa 2” asteroid explorer.)	Good

Table 11. Example outputs for the “Greece” network and the word pair “support” and “EU,” using commas and periods as delimiters

Priority	Output	Evaluation
High-frequency	*EU wa girisha shien ni yotte yuuro bouei no ketsui wo shimeshi* (EU shows a determined euro defense by supporting Greece)	OK1
Short	*EU shien kankyou totonou* (environment of complete EU support)	Bad
Ratio	*EU: girisha shien* (EU: Greece support)	Good

Table 12. MRR evaluation results (periods only)

	Good	Good /OK1	Good/ OK1/OK2
High-frequency	0.14	0.51	0.67
Short	0.29	0.36	0.58
Ratio	0.27	0.45	0.67

Table 13. Top-1 accuracy rates (periods only)

	Good	Good /OK1	Good/ OK1/OK2
High-frequency	0.08	0.36	0.50
Short	0.25	0.27	0.45
Ratio	0.17	0.32	0.52

5 Discussion

5.1 Assigning Strings to Links

Assigning character strings to the links made it possible to identify word relationships that would otherwise have been difficult to understand. However, it was not always possible to extract suitable character strings to attach to the links, because when punctuation marks were present in the character string between two words extracted by the proposed method, the character string was

Table 14. Top-5 accuracy rates (periods only)

	Good	Good /OK1	Good/ OK1/OK2
High-frequency	0.23	0.72	0.92
Short	0.37	0.48	0.75
Ratio	0.43	0.67	0.90

omitted and no character string could be extracted. This issue was particularly prevalent for word pairs that only appeared in a small number of articles. In future work, we plan to improve our approach in this respect.

5.2 Priority Equation

Next, we investigated the equations used to determine the priorities of character strings in the method using both commas and periods. We considered three priority equations: one emphasized the frequency with which the string appeared, another emphasized short character strings, and the third based the priority on the ratio between appearance frequency and length.

First, we found that the equation that focused on short character strings had the lowest performance according to all evaluation methods and criteria. We believe this was because its emphasis on short strings reduced the amount of information available to indicate the word relationships.

Next, we examine the equations emphasizing appearance frequency and the frequency/length ratio. These both produced character strings that appropriately represented the word relationships, and we found that they both yielded nearly equivalent performance when we considered output strings with extra or missing information to be correct. However, the frequency-based equation performed slightly better when only strings that properly indicated the word relationships were considered correct, and also when we allowed answers with additional information.

In addition, although the Top-1 accuracy rates for the two equations were equal, the Top-5 accuracy rate of for the frequency-based equation was higher than that of the ratio-based equation. This indicates that the frequency-based equation is likely to produce more correct answers among the top five outputs. Given that, we believe that emphasizing the occurrence frequency makes it easier to acquire character strings that appropriately indicate the word relationships.

5.3 Use of Commas and Periods

We used two methods of extracting character strings, one that only uses periods to divide the strings and another uses both commas and periods. These produced very similar evaluation results when we considered answers with additional or missing information to be correct. However, when we focused purely on answers that appropriately indicated the word relationships, the method using both commas and periods performed better. Conversely, when we allowed answers including additional information, the period-only method performed better. We believe this was because the character strings produced using periods as delimiters were longer and thus included many strings with extra information.

5.4 Discussion Comparing our Proposed Method and Other Methods

Murata et al. [9] proposed a method of using strings between two words in a sentence as the relationship of the two words in Japanese sentences. The definition sentence for an entry word “snowy moonlit night” is “a moonlit night with the presence of snow.” This shows that “snowy moonlit night” consists of two terms, “snow” and “moonlit night,” and the relationship between them is expressed using the phrase “with the presence of.” They extracted relationships such as “that is in,” “that has,” and “that was made from.”

On the other hand, in our study, a substring containing two words is extracted as representing the relationship between the two words. A substring containing two words has a wider range than a string between two words and can show the relationship between two words more clearly than a string between two words.

In the example sentence of Figure 4, “Koichi Wakata astronaut (universe flying pilot): Captain ISS, communicate with the prime minister,” the string between two words “flying” and “captain”

Table 15. Example outputs for the "Toyota" network and the word pair "Takata" and "repair," using only periods as delimiters

Priority	Output	Evaluation
High-frequency	*jidousha buhin ohte takata sei ea baggu no rikouru (kaishu/mushoushuri) mondai de, bei kain eenerugii shougyou iinkai wa mikka, jouin ni tsuzuite kouchoukai wo hiraita* (The House Energy Commerce Committee held a public hearing following the Senate on the 3rd, due to the recall (collection / free repair) of airbags manufactured by Takata, a major auto parts company)	OK1
Short	*takata sei ea baggu: 474 man dai shuuri wo* (Airbags by Takata: 4.74 million units repaired)	OK2
Ratio	*bei unyushou no douro koutsuu anzen kyoku wa 18 nichi, kekkan ga mitsukatta jidousha buhin oute takata sei ea baggu no rikouru (kaishuu/mushou shuuri) no taishou chiiki wo zenbei ni kakudai suruyou honda nado jidousha meekaa ni shiji shita to happyou shita* (The Road Traffic Safety Authority of the US Department of Transportation announced on the 18th that it had instructed automobile manufacturers such as Honda to expand the target area for the recall (collection / free repair) of Takata airbags, a major automobile part found to be defective, to the whole country)	OK1

Table 16. Example outputs for the "universe" network and the word pair "rocket" and "asteroid," using only periods as delimiters

Priority	Output	Evaluation
High-frequency	*mitsubishi juu kougyou to uchuu koukuu kenkyuu kaihatsu kikou (JAXA) wa mikka gogo, shouwakusei tansaki "hayabusa 2" wo noseta shuryoku roketto H2A26 gouki wo uchiageta* (Mitsubishi Heavy Industries and the Japan Aerospace Exploration Agency (JAXA) launched an H2A26 main rocket with the "Hayabusa 2" asteroid explorer on the afternoon of the 3rd)	OK1
Short	*kongetsu 12 gatsu ni H2A roketto de uchiagerare, shouwakusei 1999JU3 ni touchaku suru no wa 18 nen 6 gatsu* (It was launched by an H2A rocket this December, arriving at asteroid 1999JU3 on June 2018)	Bad
Ratio	*mitsubishi juu kougyou to uchuu koukuu kenkyuu kaihatsu kikou (JAXA) wa mikka gogo, shouwakusei tansaki "hayabusa 2" wo noseta shuryoku roketto H2A26 gouki wo uchiageta* (Mitsubishi Heavy Industries and the Japan Aerospace Exploration Agency (JAXA) launched an H2A26 main rocket with the "Hayabusa 2" asteroid explorer on the afternoon of the 3rd)	OK1

is only "pilot):." "pilot):" is inadequate as an expression between two words. In contrast, our method gets "Koichi Wakata astronaut (universe flying pilot): Captain ISS." "Koichi Wakata astronaut (universe flying pilot): Captain ISS" has more information than "pilot):" and shows the relationship between two words in an easy-to-understand manner. Our method is superior to the method using a string between two words.

We conducted experiments using the method based on strings between two words in a sentence. We used frequency corresponding to the use of Equation 2 in the experiments. The results are shown in Table 18. The results were lower than those in our methods. We confirmed that our methods are more effective than the method of using strings between two words.

6 Conclusions

In recent years, vast and increasing numbers of electronic texts have been posted on the Internet, so we need to find ways of automatically extracting useful information from them. Doen et

Table 17. Example outputs for the "Greece" network and the word pair "support" and "EU," using only periods as delimiters

Priority	Output	Evaluation
High-frequency	*EU wa girisha shien ni yotte yuuro bouei no ketsui wo shimeshi, kiki ga hoka no yuuro kameikoku ni tobihi suru jitai no kaihi wo mezasu* (EU shows a determined euro defense by supporting Greece, aiming to avoid a situation where the crisis extends to other euro member countries)	OK1
Short	*EU shien kankyou totonou* (environment of complete EU support)	Bad
Ratio	*EU: girisha shien goui* (EU: Greece support agreement)	Good

Table 18. Results of the use of strings between two words

	Good	Good /OK1	Good/ OK1/OK2
MRR	0.10	0.11	0.25
Top-1 accuracy rate	0.08	0.10	0.20
Top-5 accuracy rate	0.12	0.13	0.32

al. [4] proposed a method of extracting relationship information based on identifying specific keywords in the text and creating a network based on them, then constructed such a network based on the keyword "earthquake." They found that the network included nodes for unrelated concepts and proposed a method of automatically deleting them. However, their network did not include any node relationship information, making it difficult to understand these relationships.

In this study, we therefore proposed a method of extracting character strings from newspaper articles that express node relationships and assigning these strings to links in a word network. Adding character strings to links enables us to indicate the corresponding relationships.

Then, we evaluated the results produced by the proposed method in terms of the MRR, Top-1 accuracy rate, and Top-5 accuracy rate. When we considered answers with additional or missing information to be correct, the MRR was about 0.7, the Top-1 accuracy rate was about 0.6, and the Top-5 accuracy rate was about 0.9.

We also conducted experiments using two different methods of extracting character strings, one that uses commas and periods as delimiters and another that only uses periods. When we compared the performance of these two methods, we found that when we focused on answers that appropriately indicated the word relationships, the method based on both commas and periods performed better. However, when we allowed answers that included additional information, the method based on periods only was better.

Finally, when we allowed answers with additional or missing information, both methods performed very similarly.

References

1. **Bunescu, R. C., Mooney, R. J. (2005).** Subsequence kernels for relation extraction. NIPS'05, Proceedings of the 18th International Conference on Neural Information Processing Systems, pp. 171–178.
2. **Chu, E., Liu, P. (2019).** Mean sum: A neural model for unsupervised multi-document abstractive summarization. Proceedings of the 36th International Conference on Machine Learning, volume 97, pp. 1223–1232.
3. **Culotta, A., Sorensen, J. (2004).** Dependency tree kernels for relation extraction. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics ACL'04, pp. 423–429.
4. **Doen, Y., Murata, M., Otake, R., Tokuhisa, M., Ma, Q. (2014).** Construction of concept network from large numbers of texts for information examination using TF-IDF and deletion of unrelated words. Proceedings of SCIS-ISIS´14, pp. 1108–1113.
5. **Hakami, H., Bollegala, D. (2017).** Compositional approaches for representing relations between words: A comparative study. Knowledge-Based Systems, Vol. 136, pp. 172–182.

6. **Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M. (2016).** Neural relation extraction with selective attention over instances. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 2124–2133.

7. **Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Asahara, M. (1999).** Japanese Morphological Analysis System ChaSen Version 2.0 Manual.

8. **Mintz, M., Bills, S., Snow, R., Jurafsky, D. (2009).** Distant supervision for relation extraction without labeled data. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003–1011.

9. **Murata, M., Utiyama, M. (2012).** Compound word segmentation using dictionary definitions — extracting and examining of word constituent information —. ICIC Express Letters, Part B: Applications, Vol. 3, No. 3, pp. 667–672.

10. **Nenkova, A., McKeown, K. (2012).** A survey of text summarization techniques. Mining Text Data, Springer, pp. 1223–1232.

11. **Radev, D. R., Jing, H., Styś, M., Tam, D. (2004).** Centroid-based summarization of multiple documents. Information Processing and Management, Vol. 40, No. 6, pp. 919–938.

12. **Zelenko, D., Aone, C., Richardella, A. (2003).** Kernel methods for relation extraction. Journal of Machine Learning Research, Vol. 3, pp. 1083–1106.

13. **Zeng, X., He, S., Liu, K., Zhao, J. (2018).** Large scaled relation extraction with reinforcement learning. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp. 5658–5665.

14. **Zhou, G., Su, J., Zhang, J., Zhang, M. (2005).** Exploring various knowledge in relation extraction. Proceedings of the 43rd Annual Meeting of the ACL, pp. 427—434.

Article received on 24/07/2021; accepted on 13/09/2021.
Corresponding author is Masaki Murata.