Journal of Data Intelligence, Vol. 1, No. 1 (2019) 075–098
© Rinton Press

EXTRACT OF JAPANESE TEXT CHARACTERISTICS OF SIMPLIFIED CORPORA USING NON-NEGATIVE MATRIX FACTORIZATION

KOJI WAJIMA
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
kwajima@ce.slis.tsukuba.ac.jp

KEI KOGURE
Faculty of Library, Information and Media Science, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
kei.kogure@dentsu.co.jp

TOSHIHIRO FURUKAWA
Dept.Information and Computer Technology.,Science University of Tokyo
6-3-1, Niijuku, Katsushika-ku, Tokyo, 125-8585, Japan

TETSUJI SATOH
Faculty of Library, Information and Media Science, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
satoh@ce.slis.tsukuba.ac.jp

Ways of disseminating(Verbreitungsmedien) information through different media have rapidly changed owing to technological progress, especially in the field of information and communication technologies. Reflecting the changes in terms of conditions of technological progress, communication methods, and abilities have also changed. On the Internet, contents with different expressions of difficulty are mixed even though they have almost the same contents. A user who intends to search for new things or unknown things may get confused and spend a lot of time in selecting contents that are understandable for them because there are large amounts of similar contents with different difficulties. Herein, The characteristics of relevant simplified corpora are critical for everybody. In this research, we propose a method to compare two types of documents with different difficulty, and select a characteristic related to simple of expression from various characteristics related to text. In our proposed method, thousands of text characteristics are compressed and converted by Non-negative Matrix Factorization(NMF), and a basis for characterizing the simplified document is selected. The proposed method combines the characteristics of the most conducted research using the characteristics of 32 types and 2,196 dimensions. We evaluated the text characteristics in the NMF Base of the results using a classifier. As a result of applying the proposed method to two kinds of environment white papers, it became clear that an effective basis can be selected. In Addtionally, We showed estimate of the causation relationships, Optimization of the parameter. Furthermore, We showed flexibility to other media.

Keywords: NMF, LSI, LDA, Bayesian Network, Semantik, SDGs,Verbreitungsmedien

1. INTRODUCTION

Ways of disseminating (Verbreitungsmedien) information through the media have recently been changing rapidly owing to technological progress, especially in the field of information and communication technologies. Reflecting the changes in terms of conditions of technological progress, communication methods, and abilities have also changed. On the Internet, contents with different expressions of difficulty are mixed even though they have almost the same contents. A document of an easy expression called "simplified corpora" is often created for the purpose of enlightenment targeting a broad audience. However, documents that are difficult to express are created for experts in the field. A user who intends to search for new things or unknown things may get confused and spend a lot of time in selecting contents that are understandable for them because there are large amounts of similar contents with different difficulties. In this paper, we propose a method for selecting features that define a simplified corpora. Simplified corpora feature quantities compared two types of documents based on different difficulty. Simplified corpora do not require advanced reading skills; therefore, they can educate people of all ages with Internet content. Therefore, "simplified corpora" can educate to people of all ages of the readers of internet content. The characteristics of relevant simplified corpora are critical for everybody.

The following are the main contributions of this paper:

1. We use text characteristics explored in existing research without using new text characteristics and clarify features peculiar to the simplified corpora.
2. We extracted a substantial amount of characteristic data from previously conducted research and used the characteristics of 32 types and 2,196 dimensions.
3. We clarified the characteristic amount peculiar to the simplification corpus using non-negative matrix factorization (NMF) and a base selection method.

First, we extracted characteristics from different documents. Next, we performed feature conversion using NMF. In the NMF, all extracted the characteristics were used. Finally, we evaluated the NMF base using simplified corpora.

In this study, we used multivariate analysis to decompose documents into additive components. For each component, we considered a paradox of a given sentence. Therefore, the occurrence of problems decreased. The proposed method uses non-negative matrix factorization (NMF). A previous research demonstrated that NMF is superior to other multivariate analysis techniques. In addition, algorithmic expansion and improvement have been realized by the research community. Thus, NMF is regarded as the most suitable algorithm for this research.

The proposed method can be used to obtain text characteristics peculiar to the simplified corpus without using new text characteristics. We extracted a substantial amount of characteristic data from previously conducted research and used the characteristics of 32 types and 2,196 dimensions. We also evaluated all the characteristics using NMF.

In this paper, we discuss the related work in Section 2. Section 3 discusses the proposed base selection method for selecting relevant simplified corpora to facilitate understanding. Section 4 discusses the implementation and evaluation procedure that we followed, while Section 5 discusses the evaluation experiment that we conducted. Section 6 introduces the discussion. Finally, Section 7 presents a summary of the research as well as future work.

2. RELATED WORK

2.1. *Text Simplification*

In this paper, we evaluated relevant characteristics of simplified corpora. The evaluation objects are the text characteristics. Relevant characteristics of simplified corpora on previous studies, two types have been broadly classified as readability and text simplification. Research on readability evaluates the ease of understanding text characteristics. Over the past few years, several studies have addressed the research on readability. The first work on text readability was published in 1923 [1]. Determining the "topic" and "degree of difficulty" of the text in readability of text information is important. The topic of text characteristics is the purpose of the reader. When the purpose of the reader is different, text information cannot be understood. Therefore, the topic of text characteristics is important. The degree of difficulty of text information depends on the skill of the reader. When the degree of difficulty of text information is high, it cannot be understood by a beginner, and the beginner will lack motivation. Therefore, understanding the degree of difficulty of text characteristics is very important. Based on previous studies, the method that is generally used in evaluating the degree of difficulty of textual characteristics is the readability score. The characteristics applied in the previous studies on readability included "*word length*," "*total words*," and the "*hiragana ratio*"[2][3].

The object of the research on text simplification is to simplify and clarify unintelligible documents [4]. In the case of representative official languages such as English, there is little need of performing text simplification because representative official languages are prepared for simplified corpora. For instance, in the case of English Wikipedia, it is Simple English Wikipedia. On the other hand, many languages cannot use simplified corpora. The construction cost of simplified corpora is high. Therefore, text simplification is being studied for the automatic conversion for simplified corpora. Machine translation standards are used for automatic conversion of texts into simplified corpora.

Therefore, text simplification is being studied for Standards of automatic conversion into simplified corpora. The characteristics applied in the previous studies on text simplification include "*sentence length*," "*total words*," and "*the complexity of the words*" [5][6][7][8][9]. Because the proposed method can be used to evaluate the characteristics of the relevance of simplified corpora.

Therefore, this research based on research on text simplification. The following next section describes an Japanese government report on the environment. Japanese government report on the environment is evaluation object in this research.

2.2. Annual Report on the Environment

In this study, we evaluated a simplification corpus using the Japanese government's "Annual Report on the Environment"[a]. In recent years, the annual report was called the "Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan." "Annual Report on the Environment" consists of reports on the country's environmental status and its environmental conservation policies. A topic on SDGs is also shown in the latest "Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan 2018". In recent years, The international trend of social interest is sustainable development goals (SDGs) adopted by the general assembly of the United Nations. One recent SDG goal consists of 17 aims, including the following examples that are social interest in the SGDs have increased and a variety of actions have been carried out to spread [b]. The Aims relevant to the environment in SDGs are: "Clean Water and Sanitation(Goal 6)," "Affordable and Clean Energy (Goal 7)," "Responsible Consumption and Production (Goal 12)," "Climate Action (Goal 13)," "Life Below Water (Goal 14)," and "Life on Land (Goal 15)"[d].

Environmental concern is a common and important topic facing the whole world. It is a social problem next to the aging issue. In the case of "Environment Consumer Survey 2014" (in Japanese) are interested over more than 47% [e]. Superficial environmental consideration of the subject is conducted via severe correspondence. Superficial environmental consideration is called green wash[10]. Recently, the ESG component is used in evaluating corporate value. The ESG component consists of "Environment", "social", and "governance". The decision making of the investment is based on a long term corporate value. Recently year, The investment of the company is conducted by ESG investment including ESG element. The company divestment is increased when environmental consideration activity of fullness is not implemented. when environmental consideration activity of fullness is not implemented. In the company, action on environmental conservation can be implemented using a variety of environmental advertisement, environmental consideration activity, and environmental communication. Environmental advertisement is an advertisement based on an idea called "Think Globally, Act Locally." The idea was proposed by Barbara Ward. The purpose of environmental communication is to harmonize the symbiosis and information disclosure to the stakeholders. The concrete environmental communication are Company Economic Value Communication[f], Social Value Communication[g] and Cross Sector Communication[h]. In addition, consciousness, technology innovation, and the social system are important for environmental communication. "Annual Report on the Environment" is useful in understanding the social system and environmental consideration activity in the topic of the environment. However, in the "Annual Report on the Environment," technical vocabulary is used. Therefore, Understanding is difficult except for a practitioner, a researcher, a student of the specialized field. As a result, enlightenment and popularization are necessary when using text simplification.

[a] https://www.env.go.jp/en/wpaper/

[b] https://www.yoshimoto.co.jp/sdgs/

[c] https://www.ntv.co.jp/english/pressrelease/20180601.html

[d] https://www.un.org/sustainabledevelopment/

[e] Dentsu - Environment Consumer Survey 2014 (in Japanese) : https://dentsu-ho.com/articles/1011/

[f] Such as "Annual Report", "Campaign", "IR", "Product Information", etc.

[g] Such as "Public Information", "Museum", "Learning", "Proposal of Social Issue", etc.

[h] Such as "National Trust", "Partnership", etc.

Variety of actions have been undertaken to spread the “Annual Report on the Environment.” One is the “Annual Report on the Environment for Children (Kodomo Kankyo Hakusho),” whose simplified corpora are based on the “Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan.”

For this reason, in this study, we evaluate the text characteristics of the “Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan” and “Annual Report on the Environment for Children”. Our evaluation is the bases of text characteristics. In the next section and Sections 2.4 and 2.5, we explain related research that extracts text characteristics.

2.3. *Text Characteristics*

In this study, we extract a substantial amount of feature-quantity data from previously conducted studies. The proposed method outperforms a previously proposed method with respect to the quantity of the characteristic data that are analyzed. This idea is discussed in Sections 2.4 and 2.5. Our study includes Japanese text characteristics having the following properties of multiple types of characters: *hiragana*; *katakana*; *kanji*; and alphabetical, numerical, half-width, blank symbols, and double-byte characters. For this study, the number of dimensions of the text characteristics is 22. Table 1 summarizes the text characteristics with which we estimate the surface layer information.

Table 1. List of surface features of sentences

Characteristics	Dim	Frequency/Example
Number of sentences	1	[。][？][！]
Number of commas	1	[、]
Number of periods	1	[。]
Length of sentences	1	Number of bytes
Type of character	8	Hiragana, katakana [i]
Type of character (ratio)	8	(calculated [i])
Distance between commas	1	(calculated [11])
Percentage of kanji	1	(calculated [11])

[i] Unicode 10.0 Character Code Charts: https://www.unicode.org/charts/

2.4. *Topic Characteristics*

We extracted topic characteristics using an algorithm that is based on text information. The algorithm estimates the latent information of the feature extractions. Based on previous studies, two algorithms emerged: Latent Semantic Indexing (LSI) [12], which uses matrix decomposition, and Latent Dirichlet Allocation (LDA) [13], which uses a stochastic technique.

In the LSI method involving matrix decomposition, the approximation is performed using singular value decomposition(SVD). A low-rank approximation of the matrices, $\mathrm{U}^{\mathrm{T}}\mathrm{H}$, is required to minimize the total sum of the square values of a particular matrix element. LDA assumes that document w_d was created to belong to category distribution Λ using a low-rank approximation of matrices, wherein the low-rank approximation of the matrices is based on the parameters for θ and ϕ.

Herein, one row represents one document, one column represents one vocabulary word, and the element is the frequency of occurrence of vocabulary word V. The LSI and LDA schematics are presented in Figs. 1 and 2, respectively.

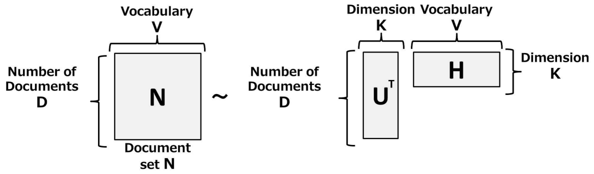


Fig. 1. Latent Semantic Indexing (LSI)

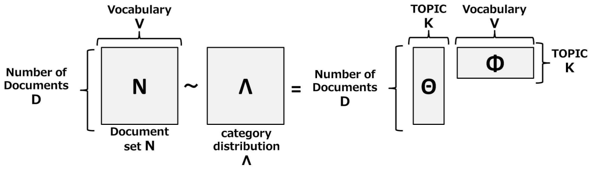


Fig. 2. Latent Dirichlet Allocation (LDA)

When LSI and LDA are employed, one row represents one document, one column represents one vocabulary word V, and element V represents the frequency of occurrence of the vocabulary word. For LSI, K represents a topic, and the elements in weighting factor matrix U denote the characteristics. For LDA, K is the topic and the elements of topic distribution θ_d denote the characteristics. In both LDA and LSI, the value of constant K is arbitrarily selected. In this paper, K was set to 300 for both LSI and LDA evaluation. Therefore, there were 600 TOPIC dimension characteristics. Table 2 summarizes the TOPIC feature quantities.

Table 2. Summary of TOPIC characteristics

Characteristics	Dim	Definition
LSI	300	Weighting factor u_D of document
LDA	300	Topic distribution θ_d of document

2.5. *Dictionary Characteristics*

Dictionaries extract the characteristics of semantic features from the text characteristics. Extensive research has been conducted on how to extract them. In this paper, we used "Semantik," which was proposed in the social systems theory of Niklase Luhmann[14]. One feature of "Semantik" is that it does not depend on the situation. This is the reason why we used a feature dictionary based on "Semantik." We compiled a feature dictionary using information obtained from previous studies. Dictionaries constitute a feature extraction method that comprises 22 feature quantities. In this study, 1,574 was the number of dimensions of the feature quantities of the dictionaries The dictionaries are composed of 22 feature quantities. Table 3 summarizes the feature quantities of the dictionaries.

Table 3. Summary of dictionary characteristics

Characteristics	Dim	Frequency/Example	
Word type [ii]	7	6519	[15]
Basic vocabulary (1) [ii]	2	697	[16]
Basic vocabulary (2) [ii]	6	6519	[16]
Basic vocabulary (3) [ii]	2	424	[16]
Semantic attributes (1) [ii]	233	697	[17]
Semantic attributes (2) [ii]	487	6519	[17]
Semantic attributes (3) [ii]	307	424	[17]
Factuality annotation [iii]	122	29262	[18]
Modality	32	32	[19]
QA end expression	38	38	[20]
IPA part of speech [iv]	14	-	[21]
Noun ratio [iv]	1	-	(calculated) [22]
MVR	1	-	(calculated) [22]
Proper noun [iv]	4	-	(calculated) [21]
NAIST JENE [v]	132	18075	[23]
Sentiment polarity (1)(2) [iii]			
(Japanese verbs)	4	5280	[24]
(Japanese nouns)	51	13314	[25]
Sentiment polarity (3) [vi]			
(Expression)	1	5234	[26]
Semantic orientations [vii]		110250	[27]
(Frequency)	2	-	(calculated) [28]
(Ratio)	2	-	(calculated) [28]
(Average)	1	-	(calculated) [28]
Fuman Category Dictionary Data (TF-IDF) [viii]	125	110,866	[29]

[ii] National Institute for Japanese Language and Linguistics
https://mmsrv.ninjal.ac.jp/bvjsl84/

[iii] Japanese FE Corpus, Japanese Sentiment Polarity Dictionary
https://www.cl.ecei.tohoku.ac.jp/index.php

[iv] ipadic version 2.7.0 : https://chasen.naist.jp/snapshot/ipadic/ipadic/doc/ipadic-ja.pdf

[v] NAIST Japanese ENE Dictionary on Wikipedia
https://github.com/masayu-a/NAIST-JENE

[vi] Evaluative Expressions : https://www.syncha.org/evaluative_expressions.html

[vii] Semantic Orientations of Words : https://www.lr.pi.titech.ac.jp/t̃akamura/pndic_ja.html

[viii] Fumankaitori Center : https://fumankaitori.com

In this paper, Among the dictionaries used, some dictionaries uses specific parts. The Semantic Orientations of Words used "*word*" and "*Japanese reading*". Further, the "National Institute for Japanese Language and Linguistics" was made for this survey. Therefore, we processed these dictionaries and show the processed point of words:

1. Remove words
   Blank symbol, [-], [0], [nado]

2. Not covered Words
   [], Include [-][〔][→][/][(][・][sonota]

3. PROPOSED METHOD

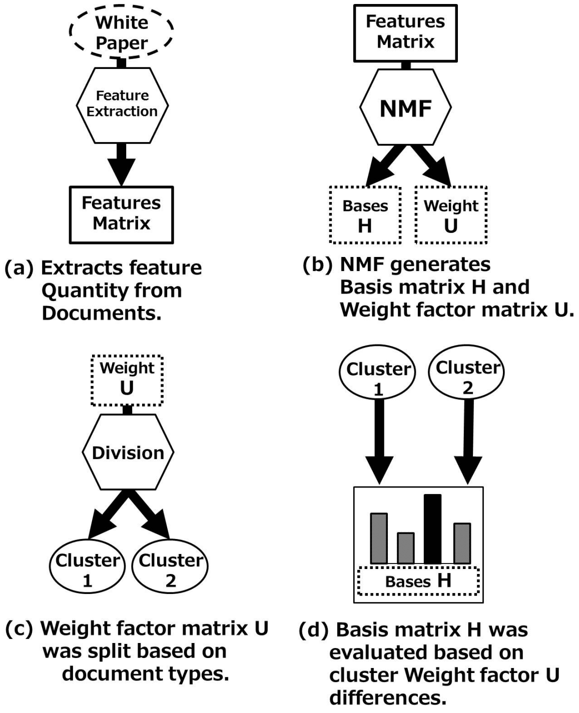
3.1. *Overview*

We propose a new method for base selection of relevant simplified corpora to simplify understanding and to disseminate enlightenment. We evaluated the following two texts: "Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan" and "Annual Report on the Environment for Children". "Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan" is gravitas report. On the other hand, the "Annual Report on the Environment for Children" is a simplified corpora based on the former text. Because simplified corpora are easily understood by all readers, they have influence. Therefore, we can distinguish simplified corpora with the characteristics from the relevance of the simplified corpora.

In this paper, we considered Gregory Bateson's definition of information: a "difference which makes a difference" [30]. From the differences observed in the "*land and map*" case, land includes the element of "*low and high*", "*building*", and "*population*". A map includes "*the map of height topography*", "*the street map*", and "*the population distribution map*" [31].

Herein, we use a map to present the result of selecting land elements, and "difference which makes a difference" based on other lands. In this paper, we propose a feature conversion and an evaluation method of the base. The proposed feature conversion uses Non-negative matrix factorization (NMF) [32]. The NMF base result can be obtained from a group of co-occurrence ingredients of the quantity of the characteristics, which is discussed in Section 3.4 Text characteristics have a contribution ratio that is based on each NMF base. Therefore, the characteristics of high contribution are based on the characteristics of the co-occurrence of the NMF base. As a consequence, the base selection result becomes a clear co-occurrence with Bateson's "difference which makes a difference."

A schematic of the proposed method is shown in Fig. 3. Figs. 3(a) to (d) show the steps involved in its procedure. Fig. 3(a) shows the feature quantities that were extracted using data presented in Sections 2.3, 2.4, and 2.5. We used the characteristics of 32 types and 2,196 dimensions to extract a substantial amount of characteristic data from previous research. Figs.3(b) to (d) show a method that can be used to process the estimates of the relevant base of the simplified corpora.



(a) Extracts feature Quantity from Documents.

(b) NMF generates Basis matrix H and Weight factor matrix U.

(c) Weight factor matrix U was split based on document types.

(d) Basis matrix H was evaluated based on cluster Weight factor U differences.

Fig. 3. Overview of proposed method

The proposed method combines the characteristics of the most conducted research using the characteristics of 32 types and 2,196 dimensions. Further, the NMF does not assume that each dimension has a correlation in observation matrix Y. Therefore, due to the correlation that may exist, NMF is not restricted like the Principal Component Analysis(PCA) or Factor Analysis(FA). An existing study has shown that NMF has a superior advantage over other existing methods[33].

We show the feature conversion method in Section 3.2, the method of splitting the NMF result in Section 3.3, and provide an evaluation method for the NMF base using weighting factors in Section 3.4.

3.2. *Feature Conversion*

Feature conversion uses NMF [32], which relies on matrix decomposition. Matrix approximation is performed using the low-rank approximation of matrices ($\mathrm{U}^{\mathrm{T}}\mathrm{H}$) to minimize the degree of divergence. In this case, observation matrix Y is presupposed to be a non-negative matrix that comprises the feature quantities in the horizontal direction. The feature quantities are generated using the data presented in Sections 2.3, 2.4, and 2.5, and the following simple NMF equation can be used to estimate the observation matrix Y:

$$Y \simeq HU \tag{1}$$

where Y denotes a rectangular observation matrix, The rectangular observation matrix Y of Eq. (2).

$$y_1, ..., y_N \subset R^{\geq 0, K} \tag{2}$$

Herein, the elements of Y and the factors of matrix elements U are the vector elements. NMF was used to convert the characteristics when the number of bases (M) was smaller than the number of dimensions (K). The numerical expression of Eq. (3) for NMF is used when $(\big(m = 1, \ldots, M\big))$ is equal to the number of bases:

$$y_n \simeq \Sigma_{m=1}^{M} h_m u_{m,n} \big(n = 1, ..., N\big) \tag{3}$$

where y_n represents he the elements of observation matrix Y, described by Eq. (3), and h_m, represents the elements of base matrix H. $u_{m,n}$ denotes the factor of the matrix elements. However, because matrix decomposition occurs within the margin of error, matrices U and H are not decided uniquely. Therefore, NMF must be optimized to minimize the margin of error. In NMF, the matrix U and base matrix H factors were initialized using random non-negative values. We alternated between the two matrices using the algorithms converged. At the point of convergence of the updated equation, the matrix U and base matrix H factors were used to generate the matrix decomposition results. The updated formula numerical formula is defined in estrangement standard. NMF defines the estrangement standard and can be used to calculate the optimum solution based on a particular estrangement standard. For the further details about updated formula, see paper [32]. We shows estrangement standard based on unified description using Parameter β described by Eq. (4).

$$D_\beta\big(y|x\big) = y\frac{y^{\beta-1} - x^{\beta-1}}{\beta - 1} - \frac{y_\beta - x_\beta}{\beta} \tag{4}$$

In Eq. (4), in the ($\beta \to 0$) case is Itakura-Saito divergence, in the ($\beta \to 1$) case is Kullback-Leibler divergence, in the ($\beta = 2$) case is square error [32]. In NMF, the optimization of the estrangement standard varies based on the created process of an observation matrix Y. The characteristics were obtained using information from dictionaries presented in previous research. The characteristics are the frequencies used, so the observation matrix is presumed to be based on the Poisson distribution. Further, the optimization of the estrangement method was accomplished using Kullback-Leibler divergence [32].

3.3. *Method of The Split*

Approximation of Y was accomplished using a low-rank approximation of matrices $\mathrm{U}^{\mathrm{T}}\mathrm{H}$, where NMF expresses Y using a linear coupling factor of matrix U and base matrix H. Eq. (5) indicates the linearity:

$$h_{j,1}\begin{pmatrix} u_{1,i} \\ u_{2,i} \\ \vdots \\ u_{m,i} \end{pmatrix} + \cdots h_{j,m}\begin{pmatrix} u_{1,i} \\ u_{2,i} \\ \vdots \\ u_{m,i} \end{pmatrix} \tag{5}$$

In NMF, the weighting factor for matrix U has a base H weight. Labeling was performed by performing binarization on the information based on the document type. We divided the data into two types of vector sets based on the label: L1 and L0. The characteristics of L1 were the simplified corpora, and those of L0 was the annual report corpora. Herein, the labeling method of division based on the label can be expressed as Eq. (6):

$$u_i = \begin{cases} x_i = 1, & u_i \in Set \quad L_1 \\ x_i = 0, & u_i \in Set \quad L_0 \end{cases} \tag{6}$$

3.4. *Base Selection*

The NMF base was evaluated using weighting factors. First, we calculated the average value of the weighting factor for NMF base m. The average value of the weighting factor was calculated for each vector set, L_1 and L_0, discussed in Section 3.3. When the weighting factor is large, the NMF base is important; however, the base of a large factor does not the basis for characterizing. In this paper, we evaluated the characteristics of base m using different weighting factor values for each vector set. If the values of the different weighting factors are large, the characterization of the base is important; if the different weighting factor value is small, this base is not important. The method for calculating the different weighting factor values is expressed in Eq. (7):

$$L_{j,m} = \frac{\Sigma_{i=1}^{N} u_{m,i}}{N} \tag{7}$$

where L_j denotes the labeling vector set and, j is 1 or 0. $L_{j,m}$ is the weighting factor value of base m in L_j. $u_{m,i}$ in Eq. (7) has base m elements of u_i from labeling vector set L_j. S is defined as the characteristics of the base set. The value method of evaluating the base m values is expressed by Eq. (8):

$$S_m = L_{1,m} - L_{0,m} \tag{8}$$

The proposed base selection method consists of converted feature quantities (3.2), divided NMF results (3.3), and base selection (3.4).

4. Implementation and Evaluation Procedure

4.1. *Experiments Environment*

Our proposed method was implemented using Python[i]. The word division and part-of-speech judgments were implemented using MeCab[j]. LSI and LDA were implemented using Gensim[k]. NMF, normalization, and L1 regularization were implemented using sci-kit-learn[l]. The observation matrix Y was created using Pandas[m] and Numpy[n]. The precision, recall, and F-measure were evaluated using scikit-learn. In LDA and LSI, a constant K value was arbitrarily selected. Based on previous studies, optimum parameter K ranged from 300 to 500 in LSI [34].

Y is a rectangular observation matrix with N rows and K columns, where the number of documents is $\left(i = 1, \ldots, N\right)$ and the number of dimensions of characteristics is $\left(j = 1, \ldots, K\right)$. Herein, the number of dimensions K consists of three feature quantities: surface layer information (22), algorithms (600), and dictionaries (1,574). Therefore, Y is an observation matrix with 68,616 rows and 2,196 columns.

The evaluation objects were the “Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan” and the “Annual Report on the Environment for Children”. Herein, “Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan” corresponds to unintelligible corpora, and “Annual Report on the Environment for Children” corresponds to simplified corpora.

We used datasets of “Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan” for 11 years between 2008 and 2018. And, We used datasets of “Annual Report on the Environment for Children” for 16 years between 2002 and 2018.

They are available as PDFs from which we [o][p] extracted text information using “pdftotext”. The text information was encoded using the UTF-8 encoding scheme. The newline characters and the loss lines of the bytes were deleted. We separated the text by the following newline characters: 「。」 In NMF, the number of documents, N was 68,616. The text information was normalized using Normalization Form KC (NFKC) [q], and the morphological analytical results were converted into canonical form. One-character words were deleted.

We evaluated the effectiveness of the proposed method using classification precision and causation estimates.

The evaluation method’s schematic is shown in Fig. 4. Fig. 4(a) classifies the evaluation documents using the characteristics of the NMF base. In Fig. 4(b), an estimate of the causation based on the document type is provided using a Bayesian network.

[i] Welcome to Python.org: https://www.python.org

[j] MeCab: https://taku910.github.io/mecab/

[k] gensim: https://radimrehurek.com/gensim/

[l] scikit-learn: https://scikit-learn.org/stable/

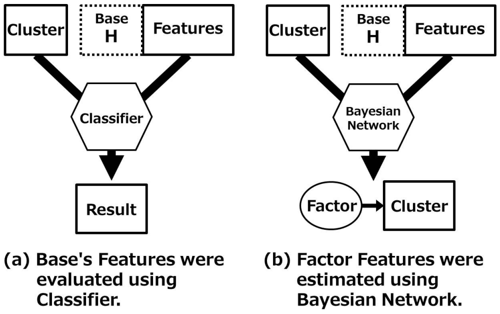
[m] Pandas: https://pandas.pydata.org/

[n] NumPy: https://www.numpy.org

[o] ISO32000-2:2017: https://www.iso.org/standard/63534.html

[p] RFC8118: https://www.rfc-editor.org/info/rfc8118

[q] Unicode Technical Reports: https://unicode.org/reports/



(a) Base's Features were evaluated using Classifier.

(b) Factor Features were estimated using Bayesian Network.

Fig. 4. Evaluation method of base

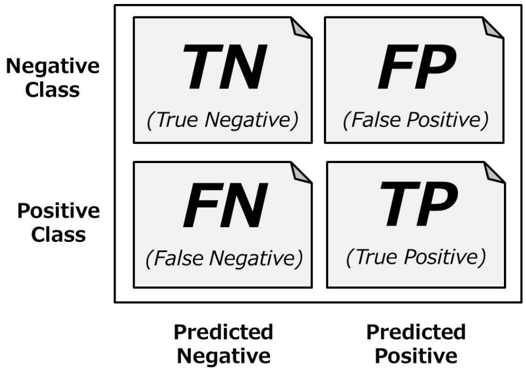
4.2. *Evaluation Indexes of Classification*

We evaluated the effectiveness of the proposed method using classification precision. The evaluation process involves the use of characteristics of the NMF base. The evaluation indices are precision, recall, and F-measure, which are represented by Eqs. (9)-(11), respectively. Fig. 5 presents a schematic of the evaluation value.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$



	Predicted Negative	Predicted Positive
Negative Class	**TN** *(True Negative)*	**FP** *(False Positive)*
Positive Class	**FN** *(False Negative)*	**TP** *(True Positive)*

Fig. 5. Contingency table

Precision denotes the accurate-answer rate of the result. Recall denotes the classification rate of accurate answers, and F-measure denotes the harmonic mean of precision and recall.

4.3. *Characteristics of Causal Relationships*

The base of the NMF consists comprised of high contribution feature quantities of the group of sets. Therefore, the characteristics of an NMF base become clear contribution characteristics. However, they do not always apply to causal relationship characteristics. Therefore, we estimated the causal relationship characteristics using the characteristics of the high-ranking contribution. The Bayesian network, which was used in evaluating the characteristics of the causal relationship [35], is based on the Bayes rule as presented in Fig. 6.

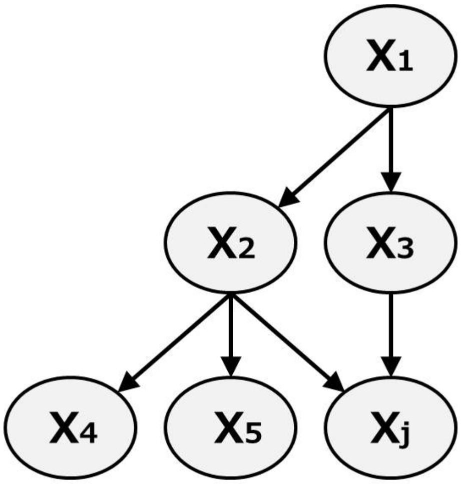


Fig. 6. Bayesian network

where X_1 denotes a parent node (i.e., a cause node), and X_2 and X_3 denote child nodes (i.e., result nodes) that become parent nodes of X_j. For probability phenomenon $P(X_1)$ and phenomenon $P(X_j)$, the probability of a result obtained from a cause is represented as $P(X_j|X_1)$, where event $P(X_1)$ denotes the cause and $P(X_j)$ denotes the result. In the Bayesian theorem, $P(X_1|X_j)$ denotes the probability that a cause creates a result, where $P(X_1)$ represents the cause and $P(X_j)$ represents the result.

Bayesian network is a model which extended the relationship between the phenomena of the cause in a large number of nodes. Each variable in a Bayesian network is a random variable. Additionally, the arrow between nodes denotes causation. Random variable and the causation are quantified as conditional probabilities and are expressed using a non-circular directed graph. From the Bayesian net work, the cause $P_\alpha(X_j)$ represents the Parent node of sets $(x_1^j, \cdots, x_i^j)$ and the child node of the result is X_j. Therefore, We can express $P(X_j|P_\alpha(X_j))$ for dependence. In such a case, the joint probability distribution of all random variable can be represented as Eq. (12).

$$P(X_1, \cdots, X_n) = \prod_{j=1}^{n} P(X_j|P_\alpha(X_j)) \tag{12}$$

When Eqs. (12) are clear, It can be expressed using the graph structure of the Bayesian network. The dependence of the random variable links between each Child node and Parent nodes. We used a contribution ratio of the quantity of characteristic of the selected base We used a contribution ratio of the quantity of characteristic of the selected base and the type of the document in the random variable $X_1, \cdots, X_n$ in constructing the Bayesian network. The dependence between variables becomes clear from the graph structure that we built. A result, we can estimate the quantity of causally associated characteristic.

5. RESULT

5.1. *Extracted the Characteristics*

The number of bases M associated with the proposed method is 50, and we evaluated the quantity of their characteristics. Fig. 7 presents the base result obtained using the proposed method. In Fig. 7, the vertical axis represents the different values of the mean of the coefficient, where the horizontal axis represents the number of bases. Therefore, the vertical-axis values are non-negative, and the base is relevant to the simplified corpora.

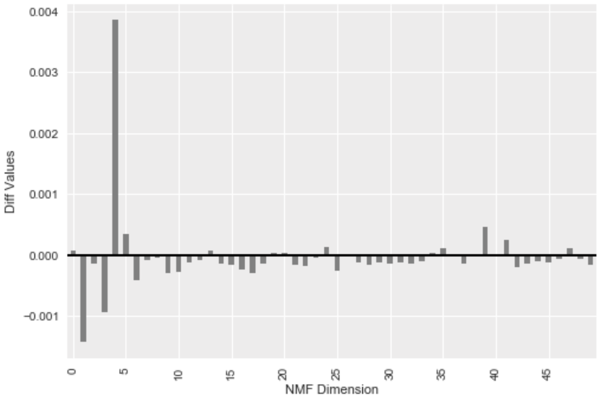


Fig. 7. Base result

From the result presented in Fig. 7, base 16 produced the biggest difference in positive value. Therefore, base 16 relates well to simplified corpora. In NMF, the factor of matrix U elements tends to become parsed [32].

5.2. *Result of Classification*

In this section, we present the result of the classification precision experiment conducted. The classification was conducted using the characteristics of the contribution of the top-ranked 100 in NMF Base 4 of Section 5.1. Table 4 summarizes the result of the classification precision.

Table 4. Result of each classification method

Classification methods	Precision	Recall	F-measure
AdaBoost	0.93	0.92	0.93
RandomForest	0.95	0.86	0.90
Multi-layer Perceptron(MLP)	0.90	0.91	0.90
K-Nearest Neighbors(K-NN)	0.91	0.86	0.88
Bagging	0.93	0.92	0.92
Gaussian Naive Bayes(GaussianNB)	0.60	0.68	0.64

All classifiers were provided with high precision, Random Forest with a precision value of 0.95 was the best classifier among the classifiers used , followed by AdaBoost and Bagging. However, only in the case of GaussianNB, the classification precision is 0.60. On the other hand, AdaBoost and Bagging were the best classifier in terms of recall with a recall value of 0.93; And, AdaBoost was the best classifier in terms of F-measure: 0.93.

5.3. *Precision Recall Curve and Receiver Operating Characteristics (ROC)*

In the results presented in Section 5.2, the F-measure value influenced the recall value. We can change the threshold value of the classifier's value. When we changed the threshold value of the classifier's value, we were shown that we could improve a reproduction rate. The model was created using training data (75%) and the classification precision was evaluated using test data (25%) The precision–recall curve, ROC, and AUC are presented in Figs. 8, and Table 5 , respectively. In the left part of Fig. 8 , the vertical axis represents recall while the horizontal axis represents precision. In the right part of Fig. 8, the vertical axis represents recall while the horizontal axis represents false positive rate.

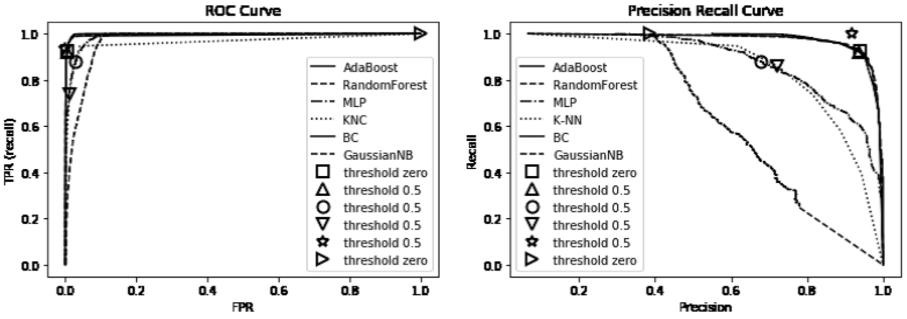


Fig. 8. Precision-Recall Curve Receiver operating characteristics (ROC)

Table 5. AUC

Classification methods	AUC
AdaBoost	0.9986
RandomForest	0.9953
Multi-layer Perceptron(MLP)	0.9894
K-Nearest Neighbors(K-NN)	0.9650
Bagging	0.9936
Gaussian Naive Bayes(GaussianNB)	0.9692

A better recall was achieved by optimizing the threshold value as shown in Figs. 8. AUCs are the evaluation indices of the classifier. In this study, all the classifiers provided good values.

5.4. *Result of Base Characteristics*

From the result obtained, the difference value of base 4 is the biggest non-negative value in Fig. 7. Therefore, base 4 is the relevant base of the simplified corpora. Table 6 summarizes the result of the features of base 4 of the proposed method. The summary presents the characteristics of the contributions of the top-ranked 20.

Table 6. Result of classification of evaluation value

Contribution of Top-Ranked 20	
Hiragana ratio	Factuality (I-conjecture-uncertain)
Semantic Orientations of Words(Negative ratio)	Factuality (B-addition)
TOPIC(LDA)	Semantic attributes (2)(4.113)
Factuality (I-judgment)	Factuality (B-solicitation)
Factuality (I-fulfilment)	Factuality (I-topic)
TOPIC(LDA)	Factuality (I-solicitation)
Part of speech(adverb)	Factuality (I-contrary connection supposition)
Factuality (B-negation)	Factuality (I-disapproval)
Factuality (I-advice)	Factuality (I-exemplification)
Sentiment polarity(Japanese verbs)(Negative（valuation）)	TOPIC(LDA)

Factuality (*) denotes the factuality annotation in Fig. 7. Hiragana ratio, factuality annotation and TOPIC (LDA) were rated high in the result of the contribution characteristics for the high-ranking of the proposed method. Hiragana ratio was the highest simplified corpora among the existing research. Therefore, the NMF base of the proposed method's result was the relevant base of the simplified corpora.

6. DISCUSSION

In section 5, We discussed the effectiveness of the proposed method. As a result, We clarified the effectiveness of proposed method. In Section 6, We show the estimate of the causation relationships, optimization of the parameters and flexibility to other media. We estimated the causation relationships using a Bayesian network. In the causation relationships, we estimated the causal relationship for the simplified corpora. In Optimization of the parameter, We demonstrated the quantity of feature choice and the number of bases.

6.1. *Result of Estimation Causal Relationships*

This section discusses the result of the characteristics of causal relationships. Table 7 summarizes the result of the causal relationship characteristics obtained using a Bayesian network. Factuality (*) denotes Factuality annotation presented in Table 7. The estimation was evaluated using the characteristics of the causal relationships of the contributions of the top-ranked 100 in the NMF Base 4 of Section . We correlated with each other with probability with more than of 0.5 from the graph structure.

Table 7. Characteristics of causal relationships

No.	Base 4
1	TOPIC(LDA)
2	Factuality (B-Invitation)
3	Factuality (B-Will)
4	Factuality (B-Appearance)
5	TOPIC(LDA)
6	Part of Speech(Verb)
7	TOPIC(LDA)
8	TOPIC(LDA)
9	Factuality (B-Repetition)
10	TOPIC(LDA)
11	Factuality (B-Completion)

NAIST JENE, Semantic attributes (2) and TOPIC(LDA) are the characteristics of the causal relationship results of the Bayesian network.

6.2. *Optimization Number of the Feature Quantity*

In this section, We show The optimization number of the feature quantity employed in the proposed method. We present a comparison result of the number of the feature quantity in table 8. The number of the bases of NMF, M is 50. The numerical value in the table denotes classification precision of the F-measure standard.

Table 8. Comparison of the number of the feature quantity(F-measure)

Classification methods	50	100	150	200	250	300
Ada Boost	0.87	0.92	0.93	0.93	0.93	0.93
Random Forest	0.87	0.91	0.91	0.91	0.90	0.88
Multi-layer Perceptron(MLP)	0.87	0.90	0.90	0.91	0.90	0.90
K-Nearest Neighbors(K-NN)	0.83	0.88	0.88	0.87	0.86	0.87
Bagging	0.89	0.92	0.92	0.92	0.93	0.92
Gaussian Naive Bayes(GaussianNB)	0.59	0.63	0.61	0.58	0.58	0.59

In table 8 , A result in most other cases, and the classification precision decreases as the selected feature quantity increases. Therefore, it is not connected directly with the improvement of the classification precision that there is much number of the select of the feature quantity. In the case of MLP and K-NN, Precision is high when there is a small number of selected feature quantity. Therefore, In this paper, the proposed method is applicable when using a small number of selected feature quantity at a given contribution ratio. Thus, the proposed method is effective.

6.3. *Compare Select Base in NMF*

In this section, Herein, we evaluate the validity of the select bases of NMF. We compare the bases of five contribution ratio high ranks. The number of bases of NMF, M is 50. We present a comparison of the number of feature quantity in table 9. The numerical value in the table represents the classification precision of the F-measure standard. The number of feature quantity is 100.

Table 9. Compare Select Base in NMF(F-measure Total)

Classification methods	4	39	5	41	24
AdaBoost	0.93	0.70	0.70	0.70	0.70
RandomForest	0.90	0.71	0.69	0.71	0.71
MLP	0.90	0.80	0.79	0.79	0.79
K-NN	0.88	0.61	0.61	0.61	0.61
Bagging	0.92	0.75	0.75	0.74	0.74
GaussianNB	0.64	0.47	0.47	0.47	0.47

When We compared the bases of five high ranks, selected the base of the proposed method yielded a result that was the highest in classification precision. Therefore, the proposed method is effective.

6.4. *Optimization Number of the NMF Parameter*

In this section, We present the optimization number of the NMF parameter K in the proposed method. We set a parameter of NMF to M = 25, 50, 100, 300 and evaluate the proposed method. We present a comparison result of the number of feature quantity in table 10. The numerical value in the table is the classification precision of the F-measure standard. Herein, the number of feature quantity is 100.

Table 10. Comparison NMF parameter K(F-measure Total)

Classification methods	25	50	100	300
AdaBoost	0.88	0.88	0.90	0.89
RandomForest	0.86	0.87	0.87	0.88
Multi-layer Perceptron(MLP)	0.89	0.89	0.83	0.80
K-Nearest Neighbors(K-NN)	0.82	0.83	0.80	0.82
Bagging	0.89	0.88	0.89	0.90
Gaussian Naive Bayes(GaussianNB)	0.63	0.64	0.56	0.52

When we compared the NMF parameter K, the difference of the classification precision based on the value of the parameter was restrictive. Thus, an evaluation experiment of K=50 is an valid evaluation experiment.

6.5. *Evaluate of Other Verbreitungsmedien*

In this section, We present the flexibility to other media. We showed flexibility to other media using Aozora Library, Research Paper, stackover flow and Internet Hakusho. The kind of the media is four kinds. We used dataset of media 1 “Aozora Bunko”[r]. Dataset of media 1 is Publications. In dataset of media 1, the number of documents, N was 14,266[36]. We used media 2 “Stack Exchange, Inc.”. Dataset of media 2 is stackover flow [s]. In dataset of media 2, the number of documents, N was 35,945[37]. We used dataset of media 3 “Japan Society for Studies in Journalism and Mass Communication”[t]. Dataset of media 3 is “Research Paper”. Research Paper is “Journal Of Mass Communication Studies”[38][39]. “Journal Of Mass Communication Studies” include predecessor of “Japanese journalism review”. We used datasets of “Japan Society for Studies in Journalism and Mass Communication” and “Japanese journalism review” for 66 years between 1952 and 2018. They are available as PDFs from which we extracted text information using “pdftotext”. We randomly selected “40, 000” documents from Research Paper. In dataset of media 3, the number of documents, N was 40,000. We used dataset of media 4 “Impress R&D”[u]. Dataset of media 4 is “Internet Hakusho”. Internet Hakusho is Annual Report on the Internet[40]. We used datasets of “Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan” for 23 years between 1996 and 2018. They are available as PDFs from which we extracted text information using “pdftotext”. In dataset of media 4, the number of documents, N was 58,866.

[r] Aozora Bunko : https://www.aozora.gr.jp

[s] Stack Exchange: Hot Questions : https://stackexchange.com/

[t] Japan Society for Studies in Journalism and Mass Communication : http://www.jmscom.org/

[u] Impress R&D : https://www.impressrd.jp/

We present a flexibility to other media from table 11 to table 16. Herein, the number of feature quantity is 100.

Table 11. Classification(AdaBoost)

Verbreitungsmedien	Precision	Recall	F-measure
Aozora Library	0.99	1.00	1.00
stackover flow	0.95	0.64	0.77
Research Paper	0.35	0.96	0.51
Internet Hakusho	0.69	0.11	0.19

Table 12. Classification(Random Forests)

Verbreitungsmedien	Precision	Recall	F-measure
Aozora Library	1.00	1.00	1.00
stackover flow	0.88	0.83	0.86
Research Paper	0.57	0.55	0.56
Internet Hakusho	0.71	0.75	0.73

Table 13. Classification(Multi-layer Perceptron)

Verbreitungsmedien	Precision	Recall	F-measure
Aozora Library	0.99	0.99	0.99
stackover flow	0.81	0.83	0.82
Research Paper	0.63	0.37	0.46
Internet Hakusho	0.65	0.81	0.72

Table 14. Classification(K-Nearest Neighbors(K-NN))

Verbreitungsmedien	Precision	Recall	F-measure
Aozora Library	0.90	0.95	0.93
stackover flow	0.74	0.73	0.74
Research Paper	0.52	0.52	0.52
Internet Hakusho	0.65	0.65	0.65

Table 15. Classification(Bagging)

Verbreitungsmedien	Precision	Recall	F-measure
Aozora Library	1.00	1.00	1.00
stackover flow	0.88	0.83	0.85
Research Paper	0.55	0.55	0.55
Internet Hakusho	0.70	0.73	0.72

Table 16. Classification(Gaussian Naive Bayes(GaussianNB))

Verbreitungsmedien	Precision	Recall	F-measure
Aozora Library	0.93	0.88	0.90
stackover flow	0.72	0.61	0.66
Research Paper	0.59	0.23	0.33
Internet Hakusho	0.57	0.85	0.68

All classifiers were provided with flexibility to other media, Random Forest with a F-measure value was the best classifier among the classifiers, followed by Bagging. Therefore, We can use the characteristic provided by proposed method for the document classification of other data sets. On the other hand, F-measure value of Research Paper of is not sufficient. Optimization of threshold of the classifiers is a future problem.

7. CONCLUSION

A new method of base selection of relevant simplified corpora was proposed. The evaluation objects were the "Annual Report on the Environment, the Sound Material-Cycle Society and Biodiversity in Japan" and "Annual Report on the Environment for Children." We combined text characteristics of existing research without using new text characteristics to clarify features peculiar to the simplified corpus. Hiragana ratio, factuality annotation, semantic attributes, and TOPIC (LDA) rated high in the contribution characteristics of the proposed method. We evaluated the characteristics in the NMF Base of the results using a classifier. The result obtained using the classifier is good, and Random Forest was the best classifier based on F-measure value obtained using a precision value of 0.94, a Recall value of 0.93, and an F-measure value of 0.93. Therefore, the proposed method is valid and effective.

Additionally, we evaluated the causal relationships and ways of optimizing the parameters.

The Bayesian network result identified the result of the causal relationship characteristics, where the feature quantities were factuality annotation, NAIST JENE, semantic attributes (2), and TOPIC(LDA). We evaluated the validity of the select bases of NMF. We compared the bases of five contribution ratio high ranks. When we compared the base of five high ranks, the select base of the proposed method produced a result with the highest in classification precision. We showed the optimization number of the feature quantity in the proposed method. In most cases showed that classification precision decreases as the number of select feature quantity increases. A result was shown, a higher number of select the feature quantity is not connected directly with the improvement of the classification precision.

We demonstrated a way of optimizing the number of NMF parameter K in the proposed method. When we compared the NMF parameter K, the difference in the classification precision based on the value of the parameter was restrictive. Therefore, there is little influence resulting from the value of the NMF parameter K. it was revealed that the proposed method of this study did not depend on the data of the evaluation experiment from a result. Additionally, We can use the characteristic provided by proposed method for the document classification of other data sets.

In addition, the proposed method is an NMF Algorithm method and based on scalar of the NMF matrix. We can perform an application to the observation matrix. Therefore, we can use the proposed method even if the dataset changes or loss occurs in the evaluation experiment. Our future work will estimate the combination of condition variables using the result of the characteristics of the proposed method.

Acknowledgements

We thank Insight Tech and National Institute of Informatics for providing us with the Fumankaitori Center - Fuman Category Dictionary Data. We are grateful to the research community for the feature quantities of the dictionaries. We are very grateful to the Japan Society for Studies in Journalism and Mass Communication and to everyone who helped make the resources of the electronic journal.

References

1. Mostafa Zamanian and Pooneh Heydari. Readability of Texts: State of the Art. *Theory & Practice in Language Studies*, 2(1), 2012.
2. J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.
3. Hideko Shibasaki. Study on Japanese Text Readability and "Easy Japanese". *Journal of Japanese Language Teaching*, 158:49–65, 2014.
4. Muhaimin Hading, Yuji Matsumoto, and Maki Sakamoto. Japanese Lexical Simplification for Non-Native Speakers. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 92–96, 2016.
5. Kajiwara Tomoyuki and Kazuhide Yamamoto. Evaluation dataset and system for Japanese lexical simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35–40, 2015.
6. Kajiwara Tomoyuki, Hiroshi Matsumoto, and Kazuhide Yamamoto. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73, 2013.
7. Tanaka Hideki, Tadashi Kumano, Isao Goto, and Hideya Mino. Rewrite Support System for Simplifying Japanese News Scripts. *Journal of Natural Language Processing*, 25(1):81–117, 2018.
8. Kajiwara Tomoyuki and Mamoru Komachi. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, 2016.
9. Suzuki Yui, Tomoyuki Kajiwara, and Mamoru Komachi. Building a non-trivial paraphrase corpus using multiple machine translation systems. In *Proceedings of ACL 2017, Student Research Workshop*, pages 36–42, 2017.
10. Mitsumasa Kotani. Development of environmental marketing and green consumerism(in japanese). *THE NAGOYA GAKUIN DAIGAKU RONSHU; Journal of Nagoya Gakuin University; SOCIAL SCIENCES*, 53(1):13–24, 2016.
11. Fukuda Makoto, Takehisa Yoshida, Makoto Yoshida, and Takefumi Kashioka. About the notation/expression of electricity fields in junior high school technology classes (in Japanese). *The Bulletin of Japanese Curriculum Research and Development*, 23(3):37–42, 2000.
12. Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.
13. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.
14. Tohru Takahashi. Societal System Differentiation and Semantik (in Japanese). *Japanese Sociological Review*, 49(4):620–634, 1998.
15. Satoru Kikuchi. University Students' Consciousness of Western Loanwords (1) (in Japanese). *The journal of the Center for Educational Research and Practices*, 4:61–73, 1994.
16. Masamitsu Sato. On Vocabulary Acquisition by Learners of Japanese:(1) Several Problems with Fundamental Vocabulary (in Japanese).
17. The National Language Research Institute (in Japanese). *A Study of The Fundamental Vocabulary For Japanese Language Teaching*, 1984.
18. Matsuyoshi Suguru, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. Factuality Annotation for Textual Information Analysis (in Japanese). *The IEICE transactions on information and systems*, 93(6):705–713, jun 2010.
19. Kazuo Fukuda. Notes on Japanese Modality(1) (in Japanese). *Universality and individuality in language*, 03(5):1–13, mar 2014.
20. Nishihara Yoko, Naohiro Matsumura, and Masahiko Yachida. Understanding of Writing Style

Patterns between Q&A in Knowledge Sharing Community (in Japanese). *Proceedings of the Annual Conference of JSAI*, 22:1–4, 2008.

21. Kudo Taku, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *In Proc. of EMNLP*, pages 230–237, 2004.
22. Keiko Nakao. A Text Analysis Based on the Part of Speech Constitution Rate (in Japanese). *Otsuma Women's University annual report. Humanities and social sciences*, 42:128–101, mar 2010.
23. Satoshi Sekine. Extended named entity ontology with attribute information. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.
24. Kobayashi Nozomi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. Collecting Evaluative Expressions for Opinion Extraction (in Japanese). *Journal of Natural Language Processing*, 12(3):203–222, 2005.
25. Higashiyama Masahiko, Kentaro Inui, and Yuji Matsumoto. Learning Sentiment of Nouns from Selectional Preferences of Verbs and Adjectives (in Japanese). *Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*, pages 584–587, 2008.
26. Kobayashi Nozomi, Kentaro Inui, and Yuji Matsumoto. Designing the Task of Opinion Extraction and Structurization (in Japanese). *IPSJ SIG Notes*, 2006(1):111–118, jan 2006.
27. Takamura Hiroya, Takashi Inui, and Manabu Okumura. Extracting Semantic Orientations of Words Using Spin Model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 133–140, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
28. Inui Takashi and Manabu Okumura. A Survey of Sentiment Analysis (in Japanese). *Journal of natural language processing*, 13(3):201–241, jul 2006.
29. Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, and Tomoya Mizumoto. Fkc corpus: a japanese corpus from new opinion survey service. *Proc. of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results*, pages 11–18, 2016.
30. Takako Kuwabara (Nakajima). The static and dynamic view point lying in the definition of information (in Japanese). *Proceedings of Annual Conference of Japan Association for Social Informatics*, 22:192–195, 2007.
31. Mamoru Ito. What is Information?: To Release the Information from a framework of Intellectualism (Special Section Article Socio-Informatics) (in Japanese). *Socio-Informatics*, 1(1):3–19, 2012.
32. Hirokazu Kameoka. Non-negative Matrix Factorization (in Japanese). *Journal of the Society of Instrument and Control Engineers*, 51(9):835–844, sep 2012.
33. Takehiko Yasukawa. Text data analysis using nonnegative matrix factorization. *Bulletin of the Computational Statistics of Japan*, 28(1):41–55, 2015.
34. Roger B. Bradford. An Empirical Study of Required Dimensionality for Large-scale Latent Semantic Indexing Applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 153–162, New York, NY, USA, 2008. ACM.
35. Yoichi Motomura. Probabilistic Reasoning using Bayesian Networks (in Japanese). *Journal of the Society of Instrument and Control Engineers*, 42(8):649–654, 2003.
36. Aozora Bunko. http://www.aozora.gr.jp(Archive date 2017-05-31).
37. Stack Exchange Data Dump.
https://archive.org/details/stackexchange(Publication date 2018-09-05).
38. JOURNAL OF MASS COMMUNICATION STUDIES.
https://www.jstage.jst.go.jp/browse/mscom/(Archive date 2019-03-25).
39. Japanese journalism review.
https://www.jstage.jst.go.jp/browse/shinbungaku/(Archive date 2019-03-25).
40. Internet Hakusho. https://iwparchives.jp/(Archive date 2019-02-11).