Interpretation Support System for Classification Patterns Using HMM in Deep Learning with Texts

Masayuki Ando
Graduate School of Engineering,
The University of Shiga Prefecture
Hikone, Japan
email: oh23mandou@ec.usp.ac.jp

Yoshinobu Kawahara
Institute of Mathematics for Industry, Kyushu University, and
RIKEN Center for Advanced Intelligence Project
Fukuoka, Japan
email: yoshinobu.kawahara@riken.jp

Wataru Sunayama
School of Engineering, The University of Shiga Prefecture
Hikone, Japan
email: sunayama.w@e.usp.ac.jp

Yuji Hatanaka
Faculty of Science and Technology, Oita University
Oita, Japan
email: hatanaka-yuji@oita-u.ac.jp

***Abstract*—This paper describes an interpretation support system for classification patterns extracted from deep learning with texts using a Hidden Markov Model (HMM), and verified its effectiveness. It is well known that classification patterns by deep learning models are often difficult to interpret the reasons derived. Therefore, an interpretation support system for deep learning classification patterns using HMMs is proposed as a tool for extracting and interpreting the learning content of deep learning. The proposed system uses an HMM to extract the contents of the learning results in deep learning with texts and provide an interface to assist in the interpretation of learned patterns. The proposed system is expected to enable system users to easily understand the complexity of deep learning, acquire new skills, and create knowledge. Verification experiments were conducted to confirm the effectiveness of the system on the basis of the learning result of deep learning classifying sentences. In the experiment, test subjects were divided into two groups. One group used the proposed system and the other used a system that displays words with high Term Frequency-Inverse Document Frequency (TFIDF) values. Both groups were instructed to provide meanings to classification patterns unusual to each output. The results show that the test subjects who used the proposed system were able to understand the meanings of the classification patterns of deep learning with texts more deeply than those who used the comparison system.**

***Index Terms*—interpretation support; deep learning, text mining; text classification; data visualization**

I. Introduction

Artificial intelligence (AI) systems based on deep learning have been rapidly increasing their number of applications. These systems have been used in a variety of fields, including image recognition, automatic driving of automobiles, automatic delivery of packages using drones, and assistance in diagnosis by doctors, with the advent of easy AI systems such as cloud AI [1] .

However, there is a black box problem in deep learning. Deep learning learns information through a very complex process and can make predictions and classifications with high accuracy. However, due to the complexity of this process, it is very difficult for humans to explain the decision criteria of deep learning.

Explainable AI (XAI) [2] has been attracting attention as a research field that focuses on explaining the reliability and fairness of deep learning models and understanding the decision criteria. The necessity of explaining what has been learned has been proposed to understand and trust the behavior of deep learning models [3] [4]. In addition, research has been conducted to try to explain the behavior of the model itself, such the model's data and correlations between variables [5], and models using counterfactual conditional statements [6]. Alternatively, there are studies that focus not only on the interpretation of model behavior but also on its stability and reliability, such as [7] [8], which evaluate model behavior by applying it to logic circuits and decision trees.

Here, if we focus on informational systems [9] [10], in which additional information is added to the output of the model where a user infers the validity or correctness of the AI's answer. In the field of image processing, for example, a method to emphasize parts of an input image that contribute to the output [11] [12] has been proposed. However, in the field of natural language processing, it is difficult to directly apply the methods used in the image processing field. In addition, just highlighting a part of the input text, known as the attention method [13], does not tell us what kind of learning is going on inside the deep learning process. In other words, the classification criteria is considered to be insufficiently explained.

Therefore, in this study, we propose a system that extracts the most likely classification patterns and assists in the interpretation of classification criteria by considering a weighted network of a trained deep learning as a single hidden Markov model (HMM), as the subject of a text classification problem. In particular, we believe that by constructing a system that enables even novice data analysts to interpret classification patterns, we can create an environment in which users of

 ISBN: 978-1-61208-870-9

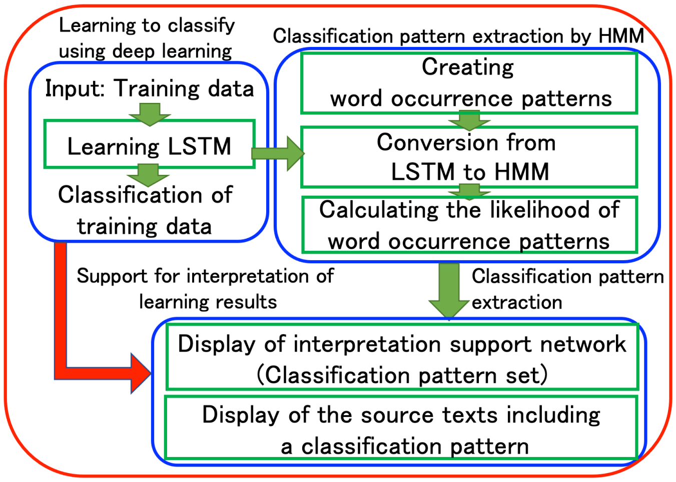

Fig. 1: System configuration

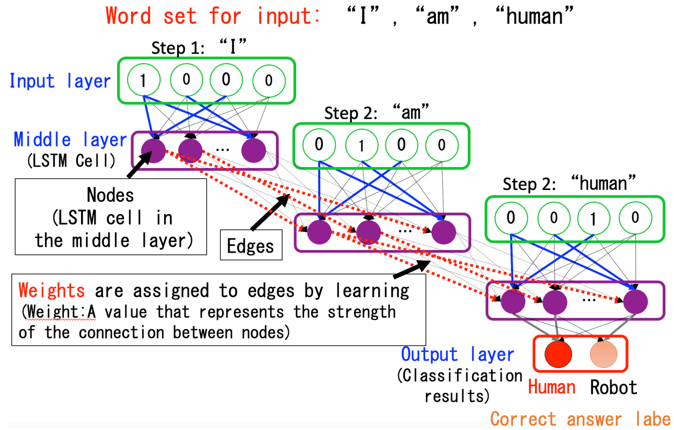

Fig. 2: Example of weighting by LSTM

cloud-based machine learning APIs and individuals who wish to perform simple text mining can easily interpret the learning results.

The following is the structure of this paper. In Section II, the structure and details of the classification pattern interpretation support system from deep learning network using HMM are described. In Section III, evaluation experiments of the proposed system are described. In Section IV, this paper is concluded.

II. Support System for Interpreting Classification Patterns Using HMM

In this section, we describe the configuration and details of our system for supporting the interpretation of classification patterns using HMMs in deep learning networks for text-based classification tasks.

A. System Configuration

The structure of the proposed system is shown in Figure 1. In the proposed system, a set of texts with correct labels is first used as training data, and classification is performed by Long Short-Term Memory (LSTM). Then, the trained weighted network is transformed into an HMM, and the likelihood of word occurrence patterns in the training text set (source text) is calculated. Finally, the word occurrence patterns with high likelihood are displayed in the interface as classification patterns, which the user can interpret.

B. Learning LSTM

In this study, we consider a deep learning model called LSTM, which is generally used to learn order patterns of time-series data, and a situation where it is applied to the problem of classifying a set of texts (Figure 2). The reason for using LSTMs is that we consider the time-series relationship of word occurrences in a text to be an important factor in this research. In addition, we do not aim to obtain high classification accuracy, but rather to build a system that encourages the interpretation of deep learning networks, which are generally difficult to interpret.

The LSTM takes as input a word vector (nouns, verbs, and adjectives in the text, with 0 and 1 representing their occurrence and non-occurrence, respectively) for each text on the basis of the set of texts given the correct answer labels. The edge weights in the LSTM with one intermediate layer are then learned so that its classification accuracy is high.

Here, to interpret the learned classification patterns, it is assumed that proper training has been performed. Therefore, we assume that the classification accuracy of the test data in the 10-fold cross-validation during training or the test data different from the training dataset is 90% or higher, and that the network does not contain substantial errors.

C. Creating Word Occurrence Patterns

To improve the interpretability of the classification patterns by making them closer to the actual text, the system uses the word occurrence patterns used in the source text for training the LSTM as input to the HMM. In this case, the word occurrence patterns to be used are all patterns that satisfy the following conditions.

- The words in a word occurrence pattern are nouns, verbs, and adjectives in the source text (adjectives may be omitted in the experiment).
- The words in a word occurrence pattern are only those in which the number of sentences where the word appears (sentence frequency) is 1% or more.
- The order of words in a word pattern should be based on the actual order in the source text.

D. Conversion of LSTM to HMM

An HMM is a non-deterministic finite state automaton model with two processes: a state and an observation symbol (output), which are a stochastic transition and output, respectively. HMMs can calculate the value (likelihood) of how plausible an observation symbol is for a given change in the state. Therefore, by applying the LSTM to the HMM, we can express the contribution of a certain word occurrence pattern to the output of the LSTM in terms of likelihood.

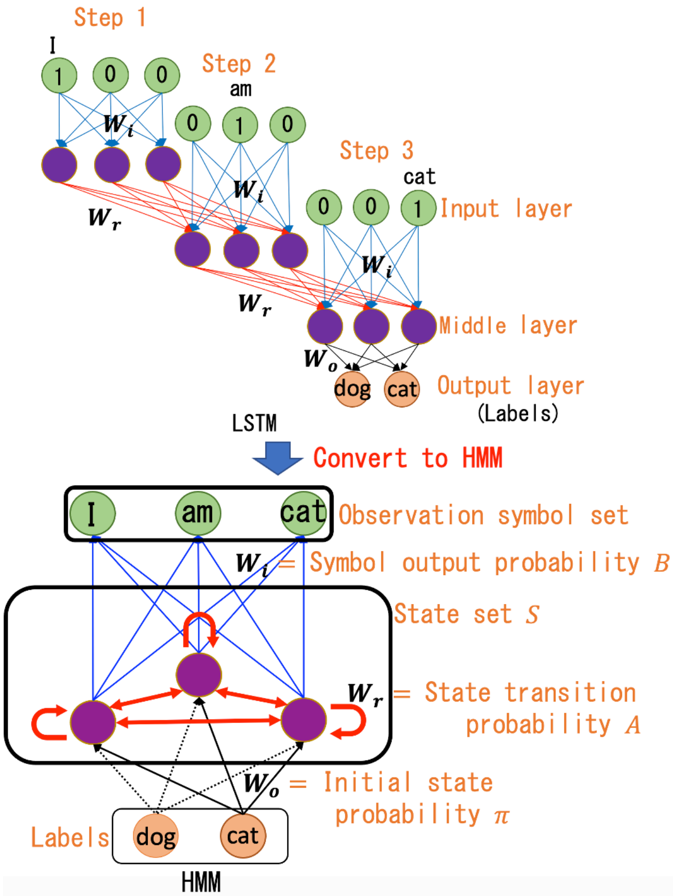

Fig. 3: Interrelationship between LSTM and HMM

In the proposed system, the weighted network obtained from the LSTM training is transformed into an HMM to estimate the likelihood of a word occurrence pattern (Figure 3). First, the input layer node of the LSTM is the observation symbol set of the HMM, and the intermediate layer node (LSTM unit) is the state set $\boldsymbol{S}$. Next, let the weight set $\boldsymbol{W_r}$ between the time series in the intermediate layer (by recursive processing) be the state transition probability $\boldsymbol{A}$, and the weight set $\boldsymbol{W_i}$ between the input layer intermediate layers be the symbol output probability $\boldsymbol{B}$. Finally, the set of weights between the output layers of the middle layer $\boldsymbol{W_o}$ is set to the initial state probability $\boldsymbol{\pi}$ (where $\boldsymbol{\pi}$ depends on the destination to be selected at that time).

E. Likelihood Estimation of Word Occurrence Patterns Using HMM

This section describes how to calculate the likelihood for the set of word occurrence patterns created in Section II-C. For the weighted network of the LSTM converted to an HMM in Section II-D, the observation series of observation symbols (the aforementioned word appearance patterns) is input to $\boldsymbol{O} = \{o_1, o_2, ...O_T\}$ (T is the length of the observation series, i.e., the length of the word appearance pattern). The number of states (number of intermediate layer nodes) is N (state number is i, j). From the aforementioned information, we can express

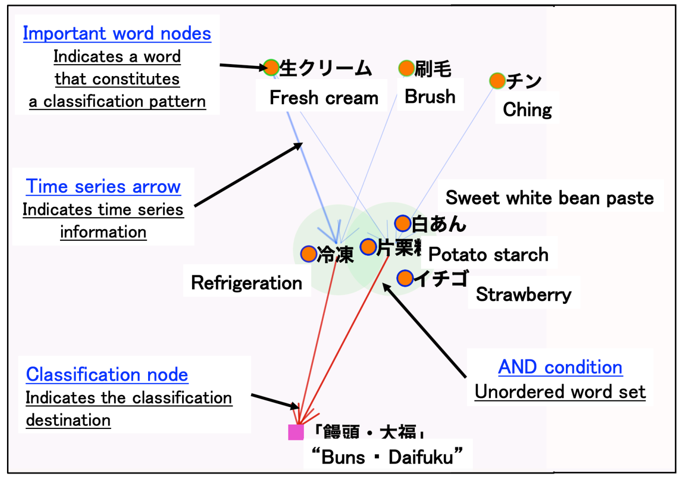

Fig. 4: Example of the system screen

the state transition probability $\boldsymbol{A}$ as (1), the symbol output probability $\boldsymbol{B}$ as (2), and the initial state probability $\boldsymbol{\pi}$ as (3).

$$\boldsymbol{A} = \{a_{ij} | a_{ij} = P(s_{t+1} = j | s_t = i)\} (1 \le i, j \le N) \quad (1)$$

$$\boldsymbol{B} = \{b_{ij}(o_t) | b_{ij}(o_t) = P(o_t | s_{t-1} = i, s_t = j)\} \\ (1 \le i, j \le N, 1 \le t \le T) \quad (2)$$

$$\boldsymbol{\pi} = \{\pi_i | \pi_i = P(s_0 = i)\} (1 \le i \le N) \quad (3)$$

Suppose that there exists a word occurrence pattern $\boldsymbol{O}$ for a classifier x. If we denote the initial state probability as $\boldsymbol{\pi}_x$, the likelihood $P(O|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B})$ is calculated by (4).

$$\begin{aligned} P(\boldsymbol{O}|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) &= \sum_{all \boldsymbol{S}} P(\boldsymbol{S}|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) P(\boldsymbol{O}|\boldsymbol{S}, \boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) \\ &= \sum_{all s_0 \dots s_T} \pi_{x s_0} a_{s_0 s_1} b_{s_0 s_1}(o_1) \cdot a_{s_1 s_2} b_{s_1 s_2}(o_2) \cdot \\ &\quad \dots \cdot a_{s_{T-1} s_T} b_{s_{T-1} s_T}(o_T) \end{aligned} \quad (4)$$

Finally, the likelihood for all word occurrence patterns are calculated by using (4), and, in the order of increasing likelihood, the word occurrence patterns are extracted as classification patterns that contribute to the classification.

F. Interpretation Support Network Display

In the interpretation support function, the set of classification patterns extracted in the previous section, which are strongly connected to the destination, is displayed as an interpretation support network. In this network, words are displayed as nodes and time-series relationship of words as arrows to make it easier to understand the words and time-series relations between words in the classification pattern. Furthermore, nodes that have arrows in both directions are

 ISBN: 978-1-61208-870-9

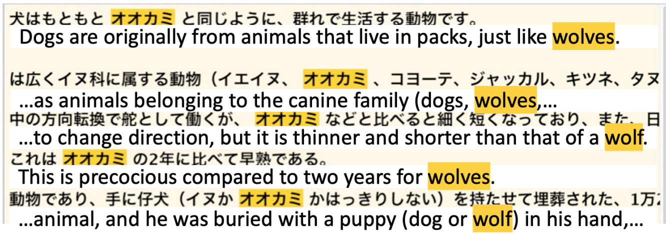

Fig. 5: Example of source text (select the word “wolf” for text about wolves)

displayed as a group. In addition, to indicate which classification pattern belongs to which destination, a red arrow connecting the destination name node and the last word node of the classification pattern is displayed.

As an example, Figure 4 shows an interpretation support network displayed from a set of texts on how to make five types of Japanese sweets [14]. The user selects a node at the bottom of the interface where the name of the classifier (in this case, “buns and daifuku”) to be interpreted is displayed. At this time, the system extracts the user’s any number of the classification patterns (in this case, five classification patterns) for the selected destination name in the order of likelihood. The extracted classification patterns are “Fresh cream $\rightarrow$ Refrigeration $\rightarrow$ Potato starch”, “Fresh cream $\rightarrow$ Potato starch $\rightarrow$ Refrigeration”, “Strawberry $\rightarrow$ Sweet white bean paste $\rightarrow$ Potato starch”, “Brush $\rightarrow$ Potato starch $\rightarrow$ Refrigeration”, and “Ching $\rightarrow$ Sweet white bean paste $\rightarrow$ Strawberry”. Then, an interpretation support network is displayed, with the words of the extracted classification patterns as nodes and the time-series relations between words as arrows. However, if there are nodes that have the same word, they will be displayed overlapping. And, if they are pointing arrows at each other, they will be displayed together as a group with no ordering relationship. Finally, by looking at the interpretation support network, the user finds patterns of what words and time-series relations between words contribute to the selected classification destination and interprets them.

G. Function for Displaying the source text of Classification Patterns

When interpreting classification patterns, it is difficult to understand the actual context in which the words were used from the word information alone. For this reason, the source text display function shows how the words in the classification pattern are actually used in the text used for training.

By selecting the user’s any words (the max is two words) on the interpretation support network, the user can see the sentence that contains the word in the source text. However, for ease of viewing, we limited the number of words displayed to ten before and after the selected word per sentence. In addition, up to two types of words can be selected, in which case, all sentences between the words are displayed. Figure 5 shows an example of the source text display of the classification pattern when the word “wolf” is selected using the text about wolves [15].

III. Verify the Effectiveness of a Text Classification Pattern Interpretation Support System Applying HMM

In this section, we describe an experiment to verify whether test subjects without deep knowledge of deep learning can interpret the classification patterns on the basis of the word occurrence patterns output by the proposed system.

A. Experimental Procedure

In the experiment, the test subjects were asked to interpret the classification patterns of the sentences classified into the “output labels” of each task for the three tasks shown in Table I. In addition, the details of the data used in each task are as follows.

Task 1 “Character dialogue”: We used 500 ”tsundere,” ”deredere,” and ”normal” character dialogues of each from ”tsundere bot,” ”deredere bot,” and ”character dialogue bot” on Twitter [16].

Task 2 “Consumer electronics reviews:” From the top 50 most popular consumer electronics products on Amazon [17]. We used 1036 “useful” (highly rated with over four stars and over ten people who said this review was helpful), “useless” (High rating, but people said this review was helpful is zero), and “low-rated” (less than two stars) reviews of each.

Task 3 “Game review,” From the top 100 most popular game software on Amazon [17]. We used 1,473 “useful,” “useless,” and “low-rated” reviews of each.

To make the interpretation easier for the test subjects and to facilitate the analysis of the interpretation results, we set an interpretation objective for each task. The experiment was conducted with 16 undergraduate and graduate students who had no deep knowledge of deep learning. They were divided into two groups: one using the proposed system and the other using the comparison system. In the group using the proposed system, the test subjects were asked to find words (one word, combinations, and time series) that contribute to classification using the proposed system. We prepared a comparison system that extracts words specific to a specified output label by the TFIDF value of (5) and presents them in a list form. The comparison system is also able to use the source text display function of the proposed system.

$$\begin{aligned}\text{TFIDF}_i \text{ of a word } i &= \text{Sentence frequency of word } i \\ &\times (\log(\frac{\text{Number of output labels}}{\text{DF value of word } i}) + 1) \qquad (5)\end{aligned}$$

The experimental procedure was interpreted by the test subjects of both groups using the following procedure. The number of classification patterns displayed by the proposed system was set to five, consisting of three words in order of increasing likelihood. The number of words displayed by the

TABLE I: EXPERIMENTAL TASKS GIVEN TO TEST SUBJECTS AND INTERPRETIVE OBJECTIVES

test subject name	Contents	Interpretive objectives
Task 1 “Character dialogue:” Output label “tsundere”	Categorization of lines of characters with specific characteristics in anime and manga: Ask test subjects to interpret the characteristics of lines of characters with “tsundere” characteristics.	Assuming you are a novelist, find a pattern of word usage specific to a “tsundere” character for your novel and give an interpretation of it.
Task 2 “Consumer electronics review:” Output label “Useful”	Categorization of reviews about popular consumer electronics on Amazon: Ask test subjects to interpret the characteristics of reviews with a large number of “this review was helpful.”	Assuming you are a reporter introducing consumer electronics, find a pattern of word usage specific to “helpful reviews” of popular consumer electronics products and give your interpretation of it.
Task 3 “Game review:” Output label “Useful”	Categorization of reviews about popular game software on Amazon: Ask test subjects to interpret the characteristics of reviews with a large number of “this review was helpful.”	Assuming that you are a reporter introducing game software, find a pattern of word usage specific to “helpful reviews” of game software and give your interpretation of it.

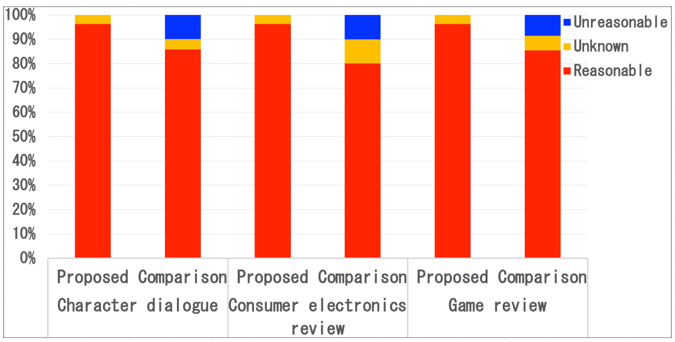

Fig. 6: Breakdown of validity of test subject’s interpretation (test subject average)

comparison system was also set to 15 in accordance with the proposed system.

Step 1 Select output labels to be interpreted: For the “Character dialogue” task, we targeted lines of characters classified as “tsundere.” For the “Consumer electronics review” and “Game review” tasks, we included reviews with a rating of four or more and a “usefulness” rating of ten or more.

Step 2 Read the “interpretation objectives” corresponding to each selected output label to understand its content.

Step 3 For the selected output, display the “interpretation support network” and find ten features (one word, combinations, chronological order, etc.) that may contribute to the output.

Step 4 Devise your own interpretations of the highlighted features in accordance with the “interpretive objectives,” using the source text display function.

B. Experimental Results and Discussion

First, the breakdown of the validity of the interpretations described by the test subjects (test subject average) is shown in Figure 6. However, the breakdown of interpretive validity was classified by one of the authors on the basis of the following definitions.

- Reasonable interpretation (reasonable): The correctness of the content can be confirmed from the source text, and it meets the “purpose of interpretation.”
- Interpretation that cannot be judged as reasonable (unknown): The intent of the content is not clear, and it cannot be judged as reasonable or not reasonable.
- Unreasonable interpretation (unreasonable): The content of the interpretation was found to be incorrect or not in line with the “purpose of the interpretation.”

Regarding the judgment of whether the subjects’ interpretations are valid or not, the author have classified them by paying attention to whether all the following criteria are met. In addition, this process was repeated at least ten times to check for errors, and the percentage of validity classified is considered to be sufficiently objective. moreover, the purpose of this experiment is only to confirm how much better the subjects can interpret the data by using the proposed system compared to the comparison system, and the comparison of the output of the systems will not be considered in this experiment.

- Checking that the features found by the subjects are included in the source text.
- Compare the subject’s interpretation with the source text, and confirm that there is at least one description that matches the claimed content.

Figure 6 shows that more than 97% of the interpretations in the proposed system were classified as reasonable interpretations, confirming their correctness. In particular, there were no interpretations that were not valid, which accounted for nearly 10% in the comparison system. In addition, the results of the proposed system showed that less than 3% of the interpretations were reasonable or unreasonable, while those of the comparison system were 5% to 10%. Therefore, we can say that the proposed system has clearer intentions and more reasonable interpretations.

Next, the percentages of source sentences that fit the interpretation (test subject average), as given by the test subjects, are shown in Figure 7. The number of source texts that match these interpretations is calculated by dividing the number of source texts for each task (500 for the “Character dialogue” task, 1036 for the “Consumer electronics review” task, and 1473 for the “Game review” task) by the percentage of source texts that match the interpretation. Regarding the number of source texts that fit the subject’s interpretation, the author counted the number of source texts that were confirmed to meet all of the following criteria. In addition, this process was

 ISBN: 978-1-61208-870-9

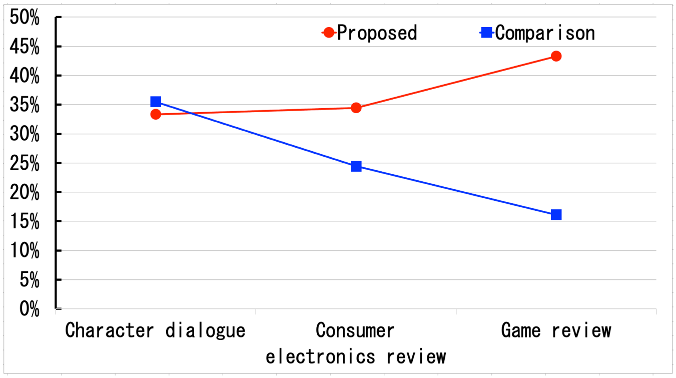

Fig. 7: Percentage of source texts to which the test subject's interpretation applies (test subject average)

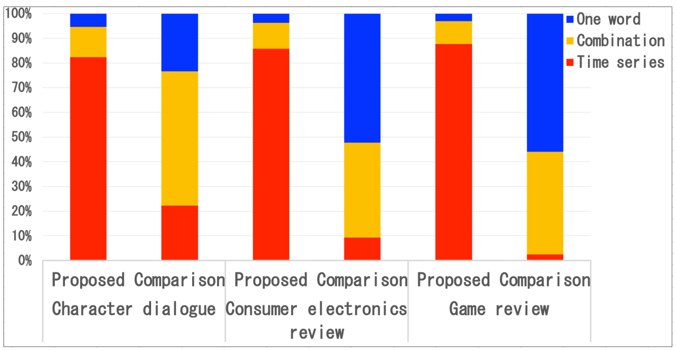

Fig. 8: Breakdown of features focused on by test subjects (test subject average)

also repeated more than 10 times to check for errors, and the number of source texts that fit the subject's interpretation is considered sufficiently objective.

- Checking that the features found by the subjects are included in the source text.
- Compare the subject's interpretation with the source text, and confirm that there is at least one description that matches the claimed content.

The results for the "Character dialogue" task were almost the same. However, those of the "Consumer electronics review" and "Game review" tasks showed that the proposed system was able to derive more interpretations that applied to the source text. In particular, for the "Game Review" task, the proposed system outperformed the results of the comparison system by nearly 30%. Therefore, we can say that the interpretation support network displayed by the proposed system was able to derive more typical interpretations that applied to a wider range of source texts.

Figure 8 shows the breakdown (test subject average) of the features (one word, combinations, time series, etc.) the test subjects focused on for interpretation. However, this breakdown was classified by one of the authors on the basis of the following definitions.

- One word: One interpretation is made from one word.
- Combination: One interpretation is made from multiple words, without any particular consideration of time-series relationships.
- Time series: One interpretation is made from multiple words, taking into account the time-series relationship.

The results in Figure 8 show that more than 80% of the interpretations in the proposed system focused on the time-series relationship of words, compared with the about 10% in the comparison system. The rest of the interpretations were based on individual words and combinations of words in the same proportion. This may be because the interpretation support network of the proposed system made the time-series relationship of the words easy to understand, and thus the test subjects paid more attention to the time-series relationship of the words and made more interpretations. However, in the comparison system, although the characteristic words of the top TFIDF were displayed, it was difficult to understand the connection between the words. Therefore, it is likely that the system was often interpreted from a single word or a combination of words with similar meanings. Therefore, we can say that the proposed system performed a typical interpretation considering the time-series relationship of words.

In summary, we confirmed that the proposed system is able to derive typical and reasonable interpretations that are applicable to a wide range of source texts with a higher rate of correct answers than the comparison system. This can be attributed to the fact that the proposed system focuses on the chronological relationship between multiple words. Furthermore, even for short texts, such as the "character dialogue" task, the proposed system is able to derive typical interpretations at the same level as referring to words with high TFIDF values.

IV. CONCLUSION AND FUTURE WORK

In this study, we proposed a classification pattern interpretation support system to classify multiple text data with an LSTM that can learn the time-series relationship of words and interpret the trained network. One of the features of this research is that it applies the network structure of the learned recursive deep learning to an HMM for processing. Therefore, the system can easily extract the time-series information of the learned features without changing the structure of the model. In the verification experiment to confirm the effectiveness of the proposed system, we confirmed that the proposed environment can result in a reasonable interpretation that covers a wide range of the original content from the classification patterns, including time-series information, even for users who are not familiar with deep learning.

In the future, we plan to investigate the effectiveness of the proposed system more objectively by statistically examining the subjects' interpretations, such as the length of their interpretations and the types of words in the sentences. In addition, we will change the input of the LSTM to a distributed representation that includes information on the relationship between words, so that the interpretation can be more focused on the meaning of the words, and also build

 ISBN: 978-1-61208-870-9

an interpretation environment for more complex deep learning networks, such as the Bidirectional Encoder Representations from Transformers (BERT). In addition, we would like to build a framework that can validate the knowledge obtained as a hypothesis by obtaining data from inside and outside the training data to support the validity of the interpretation given by the user and presenting it to the user.

REFERENCES

[1] H. Wang, et al., “Attack of the Tails: Yes, You Really Can Backdoor Federated Learning”, Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pp. 16070-16084, 2020.

[2] D. Gunning and D. Aha, “DARPA’s Explainable Artificial Intelligence (XAI) Program”, AI Magazine, Vol. 40(2), pp. 44-58, 2019.

[3] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, and F. Marcelloni, “Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?,” IEEE Computational Intelligence Magazine 14 (1), pp. 69-81, 2019.

[4] J. Haspiel, et al., “Explanations and expectations: Trust building in automated vehicles,” Companion of the ACM/IEEE International Conference on Human-Robot Interaction, ACM, pp. 119-120, 2018.

[5] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag, “Learning functional causal models with generative neural networks,” Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, pp. 39-80, 2018.

[6] R. M. J. Byrne, “Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning,” Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp. 6276-6282, 2019.

[7] G. Audemard, F. Koriche, and P. Marquis, “On Tractable XAI Queries based on Compiled Representations,” KR Proceedings 2020 Special Session on KR and Machine Learning, pp. 838-849, 2020.

[8] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, “Interpreting CNNs via decision trees,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 6261-6270, 2019.

[9] A. Barredo Arrieta, et al., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, pp. 82-115, 2020.

[10] E. Tjoa and C. Guan, “A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI,” IEEE Transactions on Neural Networks and Learning Systems 20 Oct 2020, pp. 1-21, 2020.

[11] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, and S. Behnke, “Interpretable and fine-grained visual explanations for convolutional neural networks,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9097-9107, 2019.

[12] C. Panigutti, A. Perotti, and D. Pedreschi, “Doctor XAI: an ontology-based approach to black-box sequential data classification explanations,” Proceedings of the 2020 Conference on Fairness Accountability and Transparency, pp. 629-639, 2020.

[13] M Daniluk, T Rocktaschel, J Welbl, and S Riedel, “Frustratingly Short Attention Spans in Neural Language”, In Proceedings of ICLR 2017, 2017.

[14] Cookpad Inc., “cookpad”, http://cookpad.com, <link> 2021.06.15.

[15] Wikimedia Foundation, Inc., “Wikimedia”, https://ja.wikipedia.org, <link> 2021.06.15.

[16] Twitter, Inc., “Twitter”, http://twitter.com, <link> 2021.06.15.

[17] Amazon.com, Inc., “Amazon.co.jp”, https://www.amazon.co.jp, <link> 2021.06.15.

 ISBN: 978-1-61208-870-9