Discrimination of human-written and human and machine written sentences using text consistency

Atsumu Harada
Department of Graduate School of Engineering,
Informatics Program, Kogakuin University,
Shinjuku-ku, Tokyo
em20017@ns.kogakuin.ac.jp

Danushka Bollegala
Department of Computer Science,
The University of Liverpool,
Liverpool, UK.
danushka@liverpool.ac.uk

Naiwala P. Chandrasiri
Department of Graduate School of Engineering,
Informatics Program, Kogakuin University,
Shinjuku-ku, Tokyo
chandrasiri@kogakuin.ac.jp

***Abstract*—The development of deep learning has made it possible to automatically generate sentences that could be misinterpreted as being written by humans in the field of natural language processing. As a result, the importance of the identity of the author of the text is beginning to be emphasized. In this paper, we propose a method to evaluate the consistency of sentences, which can distinguish between "sentences composed entirely of human-written texts" and "sentences with a mixture of human-written and machine-generated texts". In addition, we tested the consistency of the proposed method in an experiment, and confirmed that it was possible to discriminate two kinds of sentences in a mixed dataset of human written text and mixed text with higher accuracy than existing works. Furthermore, Kendall's rank correlation coefficient and Mann-Whitney's U-test in the sentence discrimination experiment confirmed that the proposed method showed a significant difference between the two types of sentences with a stronger correlation.**

***Index Terms*—generated text, text discrimination, text consistency, GPT-2, cosine similarity**

I. Introduction

The use of deep learning in natural language processing is expanding rapidly [1]. This is due to the development of deep learning techniques that allow for more complex tasks such as translation, advanced question answering, and sentence generation. Such a technology has made remarkable progress, and it is now possible to generate sentences that can be misinterpreted as being entirely written by humans by simply typing the beginning of a text. One of its representative deep learning language models is OpenAI's GPT-2 [2]. When this language model was first released, various problems including automatic generation of fake news was considered to occur because of its ability to generate highly accurate sentences. As a result, GPT-2 had delayed the release of the generative model. Thus, the demands for discriminating between human and machine-generated texts have risen dramatically in the modern society. However, when GPT-2 generates a large number of sentences, the content of the latter part of the text may be far from the initial content of the human input text. Table I is an example. The human written sentences and machine generated sentences using GPT-2 is shown in table I. While the first half of the content is about the album, in the second half you can see the transition of the story to black holes. It is not about grammatically wrong sentence or style, however, it shows a lack of consistency in context. Thus, machine-generated sentences tend to be less coherent than those written by humans. However, few previous studies have focused on the aspect of sentence coherence in the generation of long texts.

In this study, we proposed a new evaluation method that aims to discriminate texts that contain a mixture of human-written paragraphs and GPT-2-generated paragraphs from the perspective of sentence coherence [3] [4]. To confirm the effectiveness of the method, we conducted a total of four experiments. In the section V-A, we investigated whether the consistency of sentences could be measured using the same method as in the previous studies. In the section V-C and V-E, we measured the accuracy of the discrimination of between the two types of sentences, which is the main purpose of this study. We also confirmed that there was a significant difference in the section V-I. Finally, we measured how the accuracy of the proposed method changed under complex conditions to confirm the change in performance in actual applications in the section V-G. From those results, we confirmed that the proposed method surpassed the accuracy of the conventional method in terms of discrimination rate and gave a high evaluation value to human written sentences. In addition, there is currently a successor to the GPT-2, the more accurate GPT-3 [5]. However, since it is not open to the public, we conducted the experiments in this paper using GPT-2.

II. Related Research

A. Generative Sentence Discrimination

Current generative text discrimination methods that are using deep learning methods are popular. However, there are two issues with deep-learning detection: First, deep-learning language models take a long time to be trained. To give an

TABLE I
AN EXAMPLE OF A GENERATED PARAGRAPH USING GPT-2

Human written sentences	While not normally known for his musical talent, Elon Musk is releasing a debut album
Generated sentences	titled The Truth behind his theory of black holes. The new album was streamed nearly 1.5 million times on Spotify in just 24 hours, and was compared to Pink Floyd's The Endless River. Musk's interest in black holes began when he was 12, when his mother took him on a science lesson. "She showed us a picture of a black hole, and she said that if you threw in all the stuff that was travelling through space at the speed of light, it would just keep on travelling, forever and ever," Musk said in an interview.

example, the aforementioned GPT-3 is said to take 355 years to train a 175 billion parameter model on a single GPU[1]. Also, there is a model to detect whether the entries are written in GPT-2 [6], however this model can only be applied to GPT-2 and does not have the same generality as the present study. Furthermore, it can be seen that the improvement in accuracy requires a longer training time and the number of parameters required to improve the accuracy of the detection model increases as the number of parameters in the document generation model increase. These results suggest that detection by deep learning may be delayed when novel advanced language models are devised. In this study, we address the problem by not using deep learning itself, but only using a distributed representation of the results, which enables us to deal with sentences generated from unknown language models without learning, and at the same time, achieves faster execution speed. The second problem is that a previous study related to the discrimination of generated sentences, [7], basically requires that a single text as the input. Current sentence generation techniques produce more accurate sentences with some degree of directional specification through human inputs and hyperparameter adjustment. Therefore, when this generation technique is actually used, the beginning of a text is most likely to be written by a human being or similarly conditioned. However, conventional methods use the entire text, which may lead to false detection as the first sentence/sentences can be written by a human. Therefore, the present study is a more practical evaluation as it aims to discriminate between sentences consisting of only human-written paragraphs and mix of human-written and machine-generated paragraphs. In addition, previous studies have developed a tool that detects statistical text patterns and shows different colors for each word to help in distinguishing between human-written and machine-generated paragraphs, [8]. However, this tool only makes it easier for humans to visually understand the parts of the text that are automatically generated by the machine, and does not automatically distinguish between machine-generated paragraphs and mixed sentences. Furthermore, the target materials of this study are not only limited to news, but also novels and Wikipedia articles are included.

[1]https://lambdalabs.com/blog/demystifying-gpt-3/

B. coherence of the text

Putra et al. proposed unsupervised learning methods, PAV, SSV, and MSV, which evaluate consistency of text using graph structure [9]. PAV, which defined text coherence in terms of a formula (1), evaluated sentence coherence in terms of word overlap (uot) and sentence vector cosine similarity (cos) with the comparison of the sentences.

$$\begin{aligned}\mathrm{PAV}(S_i, S_{i-1}) &= \alpha\ \mathrm{uot}(\mathrm{Sw}_i, \mathrm{Sw}_{i-1}) \\ &\quad + (1-\alpha)\mathrm{cos}(\vec{S}_i, \vec{S}_{i-1}) \end{aligned} \tag{1}$$

The S_i in formula (1) represents the mean of the distributed representation of the words in the i-th paragraph. In addition, PAV represents the consistency of the sentence which was determined by comparison with the previous sentence. This method of evaluating consistency, and take the value of $[0, 1]$. It evaluates consistency between sentences and the larger the value, the more consistent the sentences are.

$$\mathrm{uot}(\mathrm{Sw}_i, \mathrm{Sw}_{i-1}) = \frac{|\mathrm{Sw}_i \cap \mathrm{Sw}_{i-1}|}{|\mathrm{Sw}_i \cup \mathrm{Sw}_{i-1}|} \tag{2}$$

$$\mathrm{cos}(\vec{S}_i, \vec{S}_{i-1}) = \frac{\vec{S}_{i-1}^{\mathrm{T}} \vec{S}_i}{\|\vec{S}_i\| \|\vec{S}_{i-1}\|} \tag{3}$$

In formula (1), which calculates the word overlap rate with the compared sentences in formula uot, is expressed in formula (2). For the calculation, we use the word group of the ith sentence Sw_i and the word group of the $i-1$th sentence, $\mathrm{Sw}_i - 1$. The cosine similarity of the sentence vector with the previous sentence in formula cos is expressed using the general definition formula (3). Here, we treat the distributed representation of the words averaged as the distributed representation of sentences and compute the cosine similarity between the sentences. Also, α changes the ratio of uot and cos adjusts it in the range $[0, 1]$ to determine the optimal value for each task.

$$\mathrm{tc} = \frac{1}{N} \sum_{i=2}^{N} \mathrm{sim}(S_i, S_{i-1}) \tag{4}$$

Finally, formula (4) uses the similarity between the two sentences written as sim(e.g., previous studie:PAV) and the number of sentences N to calculate the similarity between all the sentences. However, identical words are not so likely to be repeated in real sentences and are often replaced by paraphrases and pronouns. As a result, the value of the parameter α determined by learning tends to be close to zero. In addition, the best value in the insertion Task, where PAV was the best rated in a previous study, was $\alpha = 0$ [9]. These findings means that uot may not be effective in assessing the consistency of an actual text. Similarly, in the case of this study, if only the exact same words are duplicated, there can cause poor discrimination. Because deep learning does not actively paraphrase words, there are few paraphrases that are customary in English, and previous research's evaluation

method PAV records high consistency ratings for automatically generated sentences. Therefore, in this study, we propose an evaluation method that can follow the similarity of the meanings of words that are more suitable for text discrimination task by evaluating the repeat rate of this word with a different calculation formula.

III. Proposed method

In this section, we describe the two evaluation methods proposed in this study for discriminating human written text and mixed text. In the section III-A, the details of CPCO, which is an improved version of the existing method PAV is described [9]. CICO, which is an optimized version of the first method CPCO, is described in the section III-B.

In the proposed method, we used sentence-by-sentence averages of the word variances as in previous studies. Each formula consists of two consistency formulas, and the parameters α and γ can be adjusted to the optimal value for the task by changing the weights of the formulas with the value of $[0, 1]$. We also used the CoreNLP package in python for sentence/word segmentation and Lemmatization of words [10], and the pre-trained ELMo for the acquisition of distributed representations of words [11]. Also, CICO was performed under the condition that the information about the length of the human input text is known.

A. Proposed method 1:CPCO

The first proposed method, CPCO (Consistency of anti-preceding sentence using Cosine words Overlapping), is an improvement of a previous study, PAV. In CPCO, the same basic concepts of PAV are used to quantify sentence coherence in terms of sentence vector similarity and word overlap with the previous sentence. However, the Cosine similarity defined in the formula (coswot) can be used to find words of similar meaning. Equation (2) can only capture exactly same words, however, updated equation (5). coswot can capture more ambiguous expressions with similar meaning. Consistency of the sentence S_i is defined as $\mathrm{coh}(S_i)$ and to be calculated by (6). Finally, consistency across the text is defined by the equation (7).

$$\begin{aligned}&\mathrm{coswot}(\mathrm{Sw}_i, \mathrm{Sw}_{i-1}, \gamma)\\ &= \frac{|\{z \in (\mathrm{Sw}_i \times \mathrm{Sw}_{i-1}) | \cos(z) \geq \gamma\}|}{|\mathrm{Sw}_i \times \mathrm{Sw}_{i-1}|}\end{aligned} \tag{5}$$

$$\begin{aligned}\mathrm{coh}(S_i) = \alpha\ &\mathrm{coswot}(\mathrm{Sw}_i, \mathrm{Sw}_{i-1}, \gamma)\\ &+(1-\alpha)\cos(\vec{S}_i, \vec{S}_{i-1})\end{aligned} \tag{6}$$

$$\text{text coherence} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{coh}(S_i) \tag{7}$$

For the calculation of coswot, we use Sw_i and Sw_{i-1}, the words of the ith and $i-1$th sentences transformed into a distributed representation, and the threshold γ. The percentage of word pairs where the cosine similarity between words exceeds the threshold γ is expressed as $0 \sim 1$. It calculates the semantic overlap of the words in the ith and $i-1$th sentences.

B. Proposed method 2:CICO

The second proposed method, CICO (Consistency of opposing Input sentences using Cosine words Overlapping), is a more specialized version of CPCO for the problem setting of this study. It is an improvement on the PAV of the previous study and the CPCO. In the case of sentence generation where there is a human-written paragraph that is followed by a generated paragraph, both PAV and CPCO only compare the last sentence of the human-written paragraph with the first sentence of the generated paragraph explicitly assess the difference in coherence between the human-written paragraph and the generated paragraph. That is, all other consistency ratings assess the inconsistency of the generated texts, not the differences between human-written and automatically generated texts.

To address this issue, for simplicity, we have modified CICO to consider human-written sentences as a single text and to compare each sentence of the vector and generator with its entire word cloud. CICO allows us to distinguish between human-written and machine-generated sentences by passing additional information on human-written paragraphs. This makes it possible to determine whether a sentence flows naturally or not by comparing the similarity of the sentence vector to the input sentence and its vector with coswot.

The formula for CICO is defined as (8) and the average of the consistency of the whole sentence is defined by the formula (7) as well as CPCO.

$$\begin{aligned}\mathrm{coh}(S_i) = \alpha\ &\mathrm{coswot}(\mathrm{Sw}_i, \mathrm{Sw}_{\mathrm{Input}}, \gamma)\\ &+(1-\alpha)\cos(\vec{S}_i, \vec{S}_{\mathrm{Input}})\end{aligned} \tag{8}$$

In equation (8), $\mathrm{Sw}_{\mathrm{Input}}$ represents a group of distributed representations of words in the human input text, and $\vec{S}_{\mathrm{Input}}$ is the average of all sentence vectors in the human input sentences.

IV. Experimental environment

In this study, we used GPT-2 as the language model with the maximum number of parameters of 774M and other settings with the number of tokens generated by the zero-shot setting of the GPT-2 paper [2], as 300. The data set generation process takes the first sentence (120 words in average) of a paragraph written by a human and inputs it into GPT-2 as an input sentence, and generates 5 samples of sentences (about 300 words). Mixed sentences were created after that. The flow of this process is shown in Figure 1. The number of experimental datasets were randomly selected from 100 sentences in each of 7 categories of human-authored sentences, and 5 samples were generated per sentence. The resulting dataset consisted of 4200 (700 human-written, and 3500 machine-generated) texts. Note that the sentences in this dataset are about a single topic and that there are no topic transitions. In addition, the categories

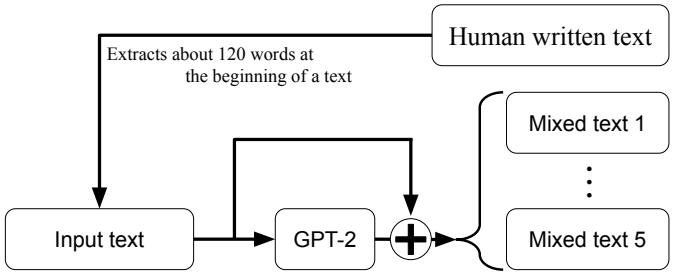

Fig. 1. How to create experimental data

of sentences used in the experiment were five categories of genres (business, entertainment, politics, sports, tech) of BBC Dataset [12], A story category was created from Gutenberg Dataset [13] and a Wikipedia category from Wikipediadumps [14]. We also plan to make the dataset publicly available on GitHub [2].

V. EXPERIMENT

A. Text consistency evaluation experiment

The evaluation formula of text consistency was updated from the previous study [9] ,text consistency evaluation experiment was conducted to confirm the effectiveness of the proposed method. In the experiments, a three-fold cross-validation method was used. Agrid search was performed in 0.1 increments within the range of $\alpha = [0, 1]$, $\gamma = [0, 1]$ to obtain the values of the parameters and evaluate the consistency. We used an insertion task, which is one of the evaluation methods of previous studies. The sentence insertion task verifies that the method is able to assess consistency by checking whether the method can record the highest consistency value when only one sentence is extracted from a text and inserted in the correct position between sentences. A previous study, performed a sentence insertion task by collecting TOEFL iBT insertion type questions [9]. However, because we wanted to assess consistency in a dataset that was not in question form, we conducted a consistency assessment experiment using the dataset we created. It only consisted of human-written paragraphs. As for the extracted sentences, we selected random sentences except for the first and last sentences. Consistency was assessed by measuring the consistency of the texts before and after inserting sentences and finding the average value.

B. Consistency Evaluation Experimental Results

The result of a consistency evaluation experiment is shown in Table II. The results show that the first proposed method, CPCO, was able to assess consistency with greater accuracy than the other methods. The second proposed method, CICO, was able to evaluate the consistency as much as the previous research method. The results confirm that both of the proposed methods work well in evaluating the consistency of the text.

[2]https://github.com/ususiop?tab=repositories

TABLE II
CONSISTENCY ASSESSMENT EXPERIMENT: ACCURACY

Method	PAV	CPCO	CICO
Accuracy	0.099	**0.112**	0.104

TABLE III
ACCURACY IN A SENTENCE DISCRIMINATION EXPERIMENT IN SENTENCES WITHOUT TOPIC TRANSITIONS

Category	PAV	CPCO	CICO
BBC_Business	0.83	**0.96**	0.95
BBC_Entertainment	0.72	0.81	**0.87**
BBC_Politics	0.72	0.88	**0.93**
BBC_Sports	0.84	0.93	**0.96**
BBC_Tech	0.83	0.92	**0.97**
Story	0.55	0.62	**0.69**
Wikipedia	0.88	0.92	**0.93**
Mean	0.767	0.865	**0.897**

C. Discrimination experiments in texts without topic transitions

As a way to evaluate the proposed method, we define two classes, first class:texts consisting only of human-written paragraphs, and the second class:texts with a mixture of human and GPT-2-generated paragraphs. We evaluated how well the proposed method can discriminate between these two classes. We consider proposed methods work well if the calculated value of consistency of a text consisting only of human-written paragraphs is higher than that of a text with a mixture of GPT-2-generated paragraphs. The experiments used a three-fold cross-validation method as well as a consistency evaluation experiment. a grid search in increments of 0.1 was performed for all methods within the range of $\alpha = [0, 1]$, $\gamma = [0, 1]$ to obtain the best parameters for the highest accuracy rate. We then calculated the percentage of correct responses for the validation data using the calculated parameter values.

D. Results of discrimination experiments without topic transitions

The averaged results of the validation data of the text discrimination experiments are listed in Table III. For PAV method, the only parameter did not change for all three partitions($\alpha = 0$) of training data showing the best accuracy rate. Although there were some variations in the suitable parameters for CPCO, the values of $\alpha = 0.7$ or $0.8, \gamma = 0$, with larger values of α and smaller values of γ, showed the best accuracy. The parameters tended to be small for both α and γ in terms of the best possible CICO accuracy, however the specific values differed from validation to validation. The results confirmed that the two proposed methods outperformed conventional PAV in all categories.

E. Discrimination experiments in texts with topic transitions

The evaluation experiment of V-C was conducted with a text consisting of one topic. However, in real-world applications, there are also texts in which the topic transitions within it

TABLE IV
ACCURACY IN A SENTENCE DISCRIMINATION EXPERIMENT IN SENTENCES WITH TOPIC TRANSITIONS

Category	PAV	CPCO	CICO
BBC_Business	0.84	**0.96**	0.89
BBC_Entertainment	0.72	0.83	**0.84**
BBC_Politics	0.72	**0.86**	**0.86**
BBC_Sports	0.88	**0.97**	0.95
BBC_Tech	0.88	**0.97**	0.93
Story	0.51	0.60	**0.63**
Wikipedia	0.88	**0.96**	0.93
Mean	0.776	**0.882**	0.862

occur. Therefore, we investigated how the discrimination performances changes for texts with topic transitions. However, we had only created a text dataset without topic transitions. Therefore, in this experiment, we created a dataset of texts with a new topic transition from the dataset IV. We randomly selected two texts from the same category in the dataset and concatenated them into a single text. As a result, we created a new dataset of 4200 topic-transitioning texts, consisting of 700 texts written by humans and 3500 mixed machine-generated texts. We also performed a three-fold cross-validation as in previous experiments and evaluated the performance using the validation text after determining the parameter values.

F. Results of Discrimination experiments in texts with topic transitions

The average of the results of the validation data for the topic transitions is listed in Table IV. The parameter values were almost all the same as in the experiment with the non-transitive text of the sectionV-C, however for CICO, only one parameter value of $\alpha = 0.1, \gamma = 0.1$ was found for the best classification rate. The results confirmed that the two proposed methods outperformed the conventional method, PAV in all categories. However, the accuracy was lower than when using a dataset of sentences with no overall topic transitions. CPCO performed best on average in sentences with topic transitions.

G. Experiment in a situation where the length of the input text is unknown

In the second proposed method, CICO, it is considered that the length of human-written input texts used to automatically generate the sentences are known. However, it is unlikely that the length of the input text is known in actual applications. Therefore, we investigated how the performance of CICO changes when the assumed length of the input texts deviates from the correct range of input texts. For the experiments, we used a dataset of sentences consisting of a single topic that was used in the discrimination experiment for sentences with no transitions in the section V-C, and we used the three-fold cross-validation method as in the past.

H. Experimental results in a situation where the length of the input text is unknown

The average percentage of correct responses for the unknown length of the input text is shown in Figure 2. The

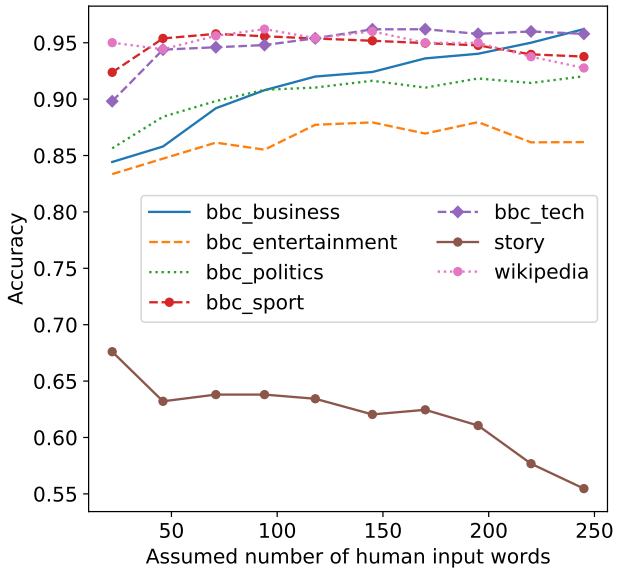

Fig. 2. Changes in the percentage of correct CICO responses with the number of assumed words in the human input text(length unknown)

parameters varied slightly with the length of the input text from the evaluation experiments with no transitions in the section V-C, however both α and γ tended to be smaller values in both cases. The results show that the accuracy tended to increase as the number of input sentences were increased except for story categories. However, some categories recorded maximum values around 120 words, which is the average number of words in the actual human-written input texts.

I. Assessment Test

We used Kendall's rank correlation coefficient and Mann-Whitney's U-test to evaluate and test whether the rating indexes correctly differentiate between sentences consisting only of human-written paragraphs and sentences with a mixture of automatically generated sentences [15] [16]. The Kendall's rank correlation coefficients were calculated setting values of 1 for human-written sentences and 0 for automatically generated sentences. This confirms the extent to which the proposed evaluation methods make a difference in the evaluation value of the text. The Mann-Whitney U test used a boxplot to visually display the distribution of rating values and to examine whether there was a statistically significant difference between texts consisting of only human-written and automatically generated ones. We used the same parameters as in the text discrimination experiment to confirm the rank order and statistical superiority of the results of the text discrimination experiment.

J. Results of Assessment Test

The results of the Kendall's rank correlation coefficients are listed in Table V. The results confirm that there is a statistically

TABLE V
LIST OF TEST RESULTS USING KENDALL'S τ
($** : p < 0.01, * : p < 0.05$)

Category	PAV	CPCO	CICO
BBC_Business	0.294**	0.439**	0.453**
BBC_Entertainment	0.213**	0.323**	0.414**
BBC_Politics	0.188**	0.388**	0.425**
BBC_Sport	0.332**	0.433**	0.456**
BBC_Tech	0.294**	0.427**	0.472**
Story	0.111*	0.069	0.100
Wikipedia	0.363**	0.420**	0.471**

significant difference for all categories except Story($** : p < 0.01, * : p < 0.05$). Both CPCO and CICO were found to have stronger correlations than previous research method, PAV. However, the previous method outperformed the proposed method in the Story category.

The boxplot for each evaluation method is shown in Figure 3. The results of the Mann-Whitney U-test showed that there were significant differences in all the evaluation methods and categories between human written texts (Raw) and sentences with a mixture of human-written and machine generated texts (Gen), respectively. Mann-Whitney's U-test between Raw-Gen and Raw-Gen, respectively, confirmed significant differences in all assessment methods and categories.

VI. DISCUSSION

From the results of the two discrimination experiments, it was confirmed that the proposed methods can discriminate human written texts and sentences with a mixture of human-written and machine generated texts with high accuracy by using the evaluated value of consistency of the sentences. In addition, we found that the method for calculating word overlap using cosine similarity proposed in this paper is more effective in the discrimination task than the conventional method used in PAV. This result is thought to be due to the fact that while the uot in PAV could only evaluate repetition of the same word, the coswot used in two proposed methods CPCO and CICO allows us to evaluate paraphrases and similar words and to capture advanced paraphrases and word transitions specific to humans.

The results of the consistency evaluation of CICO, which do not compute consistency of adjacent sentences, are more effective than the methods of previous studies. For sentences such as the dataset used in this study, in which the content of the topic is described in a few solid paragraphs, CICO was able to recognize the solidified paragraphs, and the percentage of correct responses was higher when the sentences contained similar kind of paragraphs.

Comparing the results of with and without topic transitions as a real-world application, we found that the accuracy rate of the texts of topic transitions was lower than that of no topic transitions. This is due to a decrease in the consistency rating where the topic transitions occur. However, it was confirmed that the accuracy outperformed the existing methods regardless of the presence or absence of topic transitions, confirming that it can be a practical method.

Similarly, based on the experiment a method to apply CICO to an unknown length of human input texts is to set up a tentative human input sentence length to such as about half of the input sentence. This is because the accuracy is lower with a smaller number of input sentences in categories other than story, and there is no significant decrease in accuracy even with a greater number of input sentences. In this way, we can improve the accuracy by adjusting the length of the input sentences while maintaining a higher accuracy than the previous studies.

In the test of evaluation methods, we found that the proposed method was able to express the differences between the sentences as a rating value, except for the Story category, which was similar to the sentence discrimination experiment. However, only the previous method, PAV, showed significant differences in the story category. The proposed methods also show a clearer difference between the first and third quartile boxes between Raw-Gen than previous studies, and the difference can be confirmed more clearly in CICO than in CPCO. This suggests that the proposed method can better evaluate the differences between human-written and mixed machine generated paragraphs. However, the results of the experiment as a whole showed that the evaluation values of previous and proposed methods were not very high in the Story category. In the story category sentences tend to contain more scene transitions and emotional and colloquial expressions, and simple word overlap and similarity of sentence vectors are used to evaluate the consistency of such literary expressions may not be enough. We have to address these issues in our future work.

VII. CONCLUSION

Unlike the conventional generated sentence discrimination methods, the purpose of this study was to find methods for discriminating two types of texts: human-written and mix of human and generated text. We proposed CPCO and CICO, two evaluation methods that can discriminate between above two types of text types using sentence coherence. Evaluation experiments showed that the discrimination rate of the two proposed methods outperformed the conventional method PAV. In particular, the second proposed method, CICO, was able to record clear differences in the evaluation values of sentences with a mixture of human-written and machine-generated paragraphs in a number of categories for sentences consisting of a single topic. For sentences with topic transitions, we confirmed that the first proposed method, CPCO, can discriminate sentences with higher accuracy than the other methods. All the proposed methods were found to be significantly different between the two types of text categories. In addition, CICO requires to know the length of the human-input texts. However, we have shown a method to overcome this for unknown human-input text lengths with maintaining high discrimination accuracy.

In this study, unlike the conventional generated sentence discrimination approaches, we focused on sentences that are a combination of human-written and generated paragraphs. The proposed method cannot detect correctly when all the

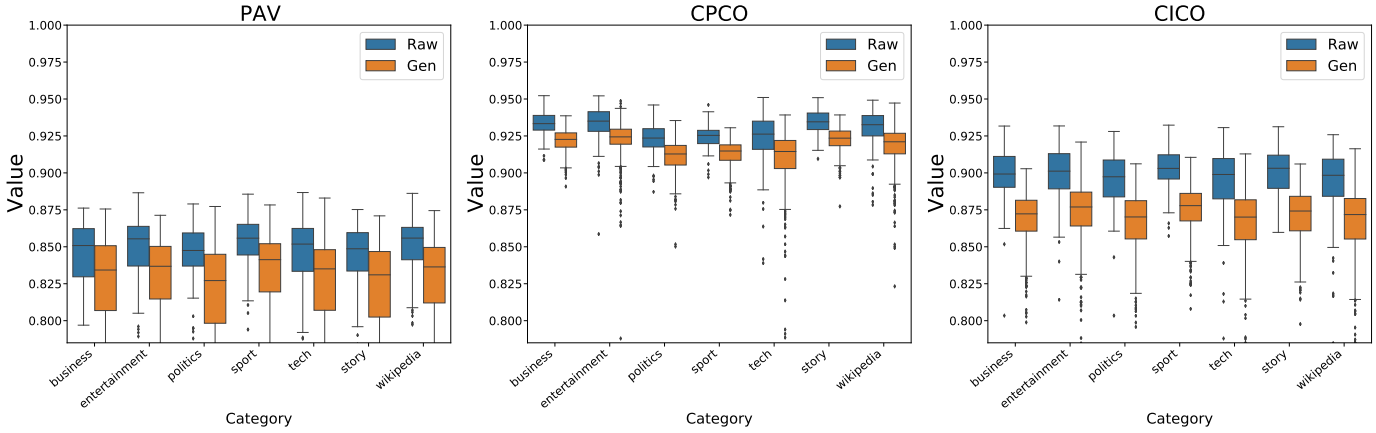

Fig. 3. Distribution by category for each evaluation method
(human written texts (Raw) and sentences with a mixtureof human-written and machine generated texts (Gen))

sentences to be discriminated are composed of generated paragraphs. Therefore, as future research, it is remained to devise an evaluation method with a value that can be used in any situation. In addition, it's an important task to improve the accuracy of CICO even when the length of the input sentence is unknown. Finally, we would like to expand this research from one-dimensional evaluation of sentences to multi-dimensional evaluation of sentences, so that we can deal with multiple text categories with higher accuracy.

REFERENCES

[1] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2020.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[3] A. Harada, D. Bollegala, and N. P. Chandrasiri, "A proposal for automatic generation of sentence detection using sentence consistency," in *NLP2020*, Mar 2020, pp. 445–448, (Japanese).

[4] A. Harada, D. Bollegala, and N. P. Chandrasiri, "A novel method for discriminating text automatically generated by machines," in *Ciber World society no.44*, Mar 2020, p. 6, (Japanese).

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[6] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, "Release strategies and the social impacts of language models," 2019.

[7] X. Zhou and R. Zafarani, "Fake news: A survey of research, detection methods, and opportunities," 2018.

[8] S. Gehrmann, H. Strobelt, and A. Rush, "GLTR: Statistical detection and visualization of generated text," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 111–116. [Online]. Available: https://www.aclweb.org/anthology/P19-3019

[9] J. W. G. Putra and T. Tokunaga, "Evaluating text coherence based on semantic similarity graph," in *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 76–85. [Online]. Available: https://www.aclweb.org/anthology/W17-2410

[10] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http:\slash\slashwww.aclweb.org\slashanthology\slashP\slashP14\slashP14-5010

[11] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202

[12] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proc. 23rd International Conference on Machine learning (ICML'06)*. ACM Press, 2006, pp. 377–384.

[13] S. Lahiri, "Complexity of Word Collocation Networks: A Preliminary Structural Analysis," in *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, April 2014, pp. 96–105. [Online]. Available: http://www.aclweb.org/anthology/E14-3011

[14] "enwiki dump progress on 20191120," (Accessed on 11/22/2019). [Online]. Available: https:\slash\slashdumps.wikimedia.org\slashenwiki\slash20191120/

[15] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938. [Online]. Available: http://www.jstor.org/stable/2332226

[16] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.