Personalized Extractive Summarization with Discourse Structure Constraints Towards Efficient and Coherent Dialog-Based News Delivery

Hiroaki Takatsu, Ryota Ando, Hiroshi Honda, Yoichi Matsuyama, and Tetsunori Kobayashi

Abstract In this paper, we propose a method to generate a personalized summary that may be of interest to each user based on the discourse structure of documents in order to deliver a certain amount of coherent and interesting information within a limited time, primarily via a spoken dialog form. We initially constructed a news article corpus with annotations of the discourse structure, users' profiles, and interests in sentences and topics. The proposed summarization model solves an integer linear programming problem with the discourse structure of each document and the total utterance time as constraints and extracts sentences that maximize the sum of the estimated degree of user's interest. The degree of interest in a sentence is estimated based on the user's profile obtained from a questionnaire and the word embeddings of BERT. Experiments confirm that the personalized summaries generated by the proposed method transmit information more efficiently than generic summaries generated based solely on the importance of sentences.

H. Takatsu (✉) · Y. Matsuyama · T. Kobayashi
Waseda University, Tokyo, Japan
e-mail: takatsu@pcl.cs.waseda.ac.jp

Y. Matsuyama
e-mail: matsuyama@pcl.cs.waseda.ac.jp

T. Kobayashi
e-mail: koba@waseda.jp

R. Ando
Naigai Pressclipping Bureau, Ltd., Tokyo, Japan
e-mail: ando@naigaipc.co.jp

H. Honda
Honda Motor Co., Ltd., Tokyo, Japan
e-mail: hiroshi_01_honda@jp.honda

S. Stoyanchev et al. (eds.), *Conversational AI for Natural Human-Centric Interaction*,
Lecture Notes in Electrical Engineering 943,
https://doi.org/10.1007/978-981-19-5538-9_4

1 Introduction

As people's interests and preferences diversify, the demand for personalized summarization technology has increased [1]. Summaries are classified as generic or user-focused, based on whether they are specific to a particular user [2]. Unlike generic summaries generated by extracting important information from the text, user-focused summaries are generated based not only on important information but also on a user's interests and preferences. Most user-focused summarization methods rank sentences based on a score calculated considering user's characteristics and subsequently generate a summary by extracting higher-ranked sentences [3–5]. However, such conventional user-focused methods tend to generate incoherent summaries. Generic summarization methods, which consider the discourse structure of documents, have been proposed to maintain coherence [6–8]. To achieve both personalization and coherence simultaneously, we propose a method to extract sentences that may be of interest according to a user's profile and generate a personalized summary for each user while maintaining coherence based on the discourse structure of documents.

As mobile personal assistants and smart speakers become ubiquitous, the demand for spoken dialog technology has increased. However, dialog-based media is more restrictive than textual media. For example, when listening to an ordinary smart speaker, users can not skip unnecessary information or skim only for necessary information. Thus, it is crucial for future dialog-based media to extract and efficiently transmit information that the users are particularly interested in without excess or deficiencies.

We utilize the proposed personalized summarization method for a spoken dialog system that delivers news as a realistic application [9]. This news dialog system proceeds the dialog according to a primary plan to explain the summary of the news article and subsidiary plans to transmit supplementary information through question answering. As long as the user is listening passively, the system transmits the content of the primary plan. The personalized primary plan generation problem can be formulated as follows:

> From N documents with different topics, sentences that may be of interest to the user are extracted based on the discourse structure of each document. Then the contents are transmitted by voice within T seconds.

Specifically, this problem can be formulated as an integer linear programming problem, which extracts sentences that maximize the sum of the degree of user's interest in the sentences of each document with the discourse structure of documents and the total utterance time T as constraints. The degree of interest in a sentence is estimated based on the user's profile obtained from a questionnaire and the word embeddings of bidirectional encoder representations from transformers (BERT) [10]. To evaluate the effectiveness of the proposed method, we construct a news article corpus with annotations of the discourse structure, users' profiles, and interests in sentences and topics.

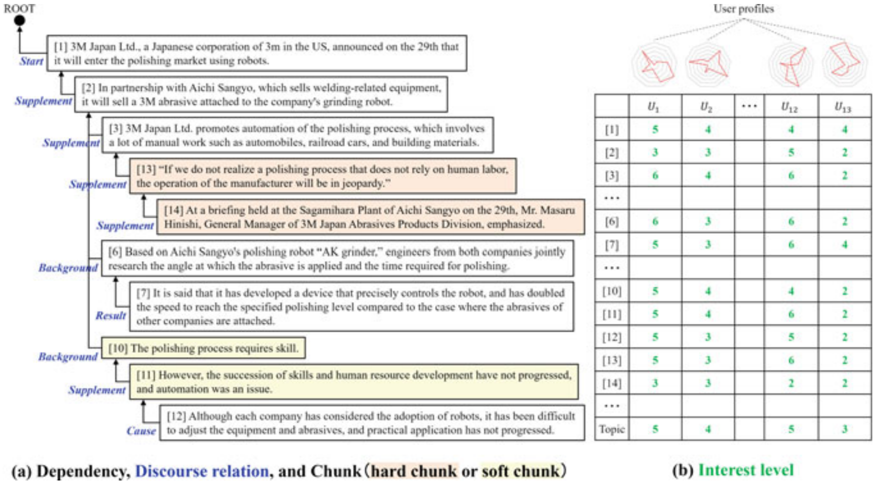
The rest of this paper is organized as follows. Section 2 overviews the discourse structure annotation and interest data collection. Section 3 describes the proposed method. Section 4 evaluates its performance. Section 5 provides the conclusions and future prospects.

2 Datasets

We constructed a news article corpus with annotations of the discourse structure, users' profiles, and interests in sentences and topics. Figure 1 shows an example of the annotation results. Experts annotated the inter-sentence dependencies, discourse relations, and chunks for the Japanese news articles. The users' profiles and interests in the sentences and topics of news articles were collected via crowdsourcing.

2.1 Discourse Structure Dataset

Two web news clipping experts annotated the dependencies, discourse relations, and chunks for 1,200 Japanese news articles. Each article contained between 15–25 sentences. The articles were divided into six genres: sports, technology, economy, international, society, and local. In each genre, we manually selected 200 articles to minimize topic overlap. The annotation work was performed in the order of depen-



[1] 3M Japan Ltd., a Japanese corporation of 3m in the US, announced on the 29th that it will enter the polishing market using robots.

[2] In partnership with Aichi Sangyo, which sells welding-related equipment, it will sell a 3M abrasive attached to the company's grinding robot.

[3] 3M Japan Ltd. promotes automation of the polishing process, which involves a lot of manual work such as automobiles, railroad cars, and building materials.

[13] "If we do not realize a polishing process that does not rely on human labor, the operation of the manufacturer will be in jeopardy."

[14] At a briefing held at the Sagamihara Plant of Aichi Sangyo on the 29th, Mr. Masaru Hinishi, General Manager of 3M Japan Abrasives Products Division, emphasized.

[6] Based on Aichi Sangyo's polishing robot "AK grinder," engineers from both companies jointly research the angle at which the abrasive is applied and the time required for polishing.

[7] It is said that it has developed a device that precisely controls the robot, and has doubled the speed to reach the specified polishing level compared to the case where the abrasives of other companies are attached.

[10] The polishing process requires skill.

[11] However, the succession of skills and human resource development have not progressed, and automation was an issue.

[12] Although each company has considered the adoption of robots, it has been difficult to adjust the equipment and abrasives, and practical application has not progressed.

	U_1	U_2	…	U_{12}	U_{13}
[1]	5	4		4	4
[2]	3	3		5	2
[3]	6	4		6	2
…					
[6]	6	3		6	2
[7]	5	3		6	4
…					
[10]	5	4		4	2
[11]	5	4		6	2
[12]	5	3		5	2
[13]	5	3		6	2
[14]	3	3		2	2
…					
Topic	5	4		5	3

(a) Dependency, Discourse relation, and Chunk (hard chunk or soft chunk)

(b) Interest level

Fig. 1 Example of the annotation results

dencies, discourse relations, and chunks. The discourse unit was a sentence, which represents a character string separated by an ideographic full stop.

2.1.1 Dependency Annotation

The conditions in which sentence j can be specified as the parent of sentence i are as follows:

- In the original text, sentence j appears before sentence i.
- The flow of the story is natural when reading from the root node in order according to the tree structure and reading sentence i after sentence j.
- The information from the root node to sentence j is the minimum information necessary to understand sentence i.
- If it is possible to start reading from sentence i, the parent of sentence i is the root node.

2.1.2 Discourse Relation Annotation

A discourse relation classifies the type of semantic relationship between the child sentence and the parent sentence. We defined the following as discourse relations: *Start*, *Result*, *Cause*, *Background*, *Correspondence*, *Contrast*, *Topic Change*, *Example*, *Conclusion*, and *Supplement*. An annotation judgment was made while confirming whether both the definition of the discourse relation and the dialog criterion were met. The dialog criterion is a judgment based on whether the response is natural according to the discourse relation. For example, the annotators checked whether it was appropriate to present a child sentence as an answer to a question asking the cause, such as "Why?" after the parent sentence.

2.1.3 Chunk Annotation

A chunk is a highly cohesive set of sentences. If a parent sentence should be presented with a child sentence, it is regarded as a chunk.

A *hard chunk* occurs when the child sentence provides information essential to understand the content of the parent sentence. Examples include when the parent sentence contains a comment and the child sentence contains the speaker's information or when a procedure is explained over multiple sentences.

A *soft chunk* occurs when the child sentence is useful to prevent a biased understanding of the content of the parent sentence, although it does not necessarily contain essential information to understand the parent sentence itself. An example is explaining the situation in two countries related to a subject, where the parent sentence contains one explanation and the child sentence contains another.

2.1.4 Annotation Quality

A one-month training period was established, and discussions were held until the annotation criteria of the two annotators matched. To validate the inter-rater reliability, the two annotators annotated the same 34 articles after the training period. The Cohen's kappa of dependencies, discourse relations, and chunks were 0.960, 0.943, and 0.895, respectively. To calculate kappa of the discourse relations, the comparison was limited to the inter-sentence dependencies in which the parent sentence matched. To calculate kappa of the chunks, we set the label of the sentence selected as the hard chunk, soft chunk, and other to "1, 2, and 0," respectively. Then we compared the labels between sentences. Given the high inter-rater reliability, we concluded that the two annotators could cover different assignments separately.

2.2 Interest Dataset

Participants were recruited via crowdsourcing. They were asked to answer a profile questionnaire and an interest questionnaire. We used 1,200 news articles, which were the same as those used in the discourse structure dataset. We collected the questionnaire results of 2,507 participants. Each participant received six articles, one from each genre. The six articles were distributed so that the total number of sentences was as even as possible across participants. Each article was reviewed by at least 11 participants.

2.2.1 Profile Questionnaire

The profile questionnaire collected the following information: gender, age, residential prefecture, occupation type, industry type, hobbies, frequency of checking news (daily, 4–6 days a week, 1–3 days a week, or 0 days a week), typical time of day news is checked (morning, afternoon, early evening, or night), methods to access the news (video, audio, or text), tools used to check the news (TV, newspaper, smartphone, etc.), newspapers, websites, and applications used to check the news (Nihon Keizai Shimbun, LINE NEWS, SNS, etc.), whether a fee was paid to check the news, news genre actively checked (economy, sports, etc.), and the degree of interest in each news genre (not interested at all, not interested, not interested if anything, interested if anything, interested, or very interested).

2.2.2 Interest Questionnaire

After reading the text of the news article, participants indicated their degree of interest in the content of each sentence. Finally, they indicated their degree of interest in the topic of the article. The degree of interest was indicated on six levels: 1, not interested at all; 2, not interested; 3, not interested if anything; 4, interested if anything; 5, interested; or 6, very interested.

3 Methods

3.1 Inter-Sentence Dependency Parsing

Figure 2 schematically diagrams the proposed model. First, the sentences are inputted with the title added as ROOT: $S = (s_0 = \text{ROOT}, s_1, \ldots, s_n)$. The words of the sentence are given to BERT to generate the embedding of the top layer corresponding to the [CLS] token. Next, the sentence vector and the embedding of the auxiliary features of the sentence are concatenated and given to the bidirectional model [11] of the gated recurrent unit (GRU) [12]. Here, the auxiliary features are the sentence and paragraph positions in the document and the sentence position in the paragraph. $\boldsymbol{h}_i$ is a vector that concatenated the outputs of the hidden layers in the forward and backward directions of the GRU corresponding to the i-th sentence. Based on the head selection model [13], the probability $P_{head}\left(s_j|s_i, S\right)$ that s_j is the parent of s_i is calculated as

$$P_{head}\left(s_j|s_i, S\right) = \frac{\exp\left(g\left(\boldsymbol{h}_j, \boldsymbol{h}_i\right)\right)}{\sum_{k=0}^{N} \exp\left(g\left(\boldsymbol{h}_k, \boldsymbol{h}_i\right)\right)} \quad (1)$$

$$g\left(\boldsymbol{h}_j, \boldsymbol{h}_i\right) = \boldsymbol{v}_h^{\mathsf{T}} \tanh\left(\boldsymbol{U}_h \boldsymbol{h}_j + \boldsymbol{W}_h \boldsymbol{h}_i\right) \quad (2)$$

where $\boldsymbol{v}_h$, $\boldsymbol{U}_h$, $\boldsymbol{W}_h$ are weight parameters.

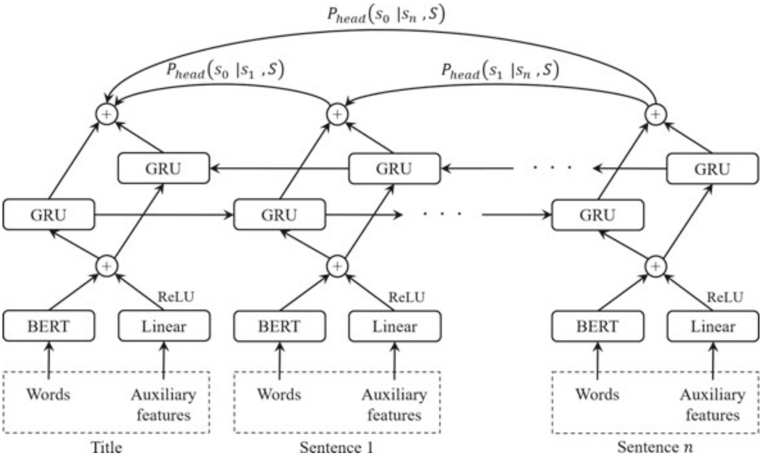


Fig. 2 Inter-sentence dependency parser

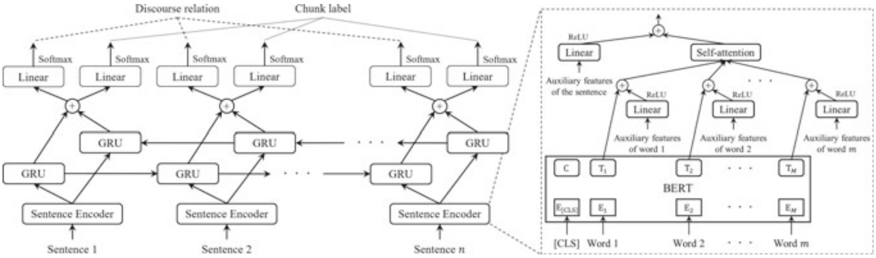


Fig. 3 Discourse relation and chunk estimator

3.2 Discourse Relation Classification and Chunk Detection

The discourse dependency tree is decomposed into sentence sequences from the root node to the leaf nodes (hereafter, root-to-leaf). Discourse relations and chunk labels are estimated by sequence labeling for the root-to-leaf sentences. Figure 3 schematically diagrams the proposed model. The total loss function $\mathcal{L}_{total}$ of multi-task learning is defined by the weighted sum of the loss function $\mathcal{L}_r$ of the discourse relation classification task and the loss function $\mathcal{L}_c$ of the chunk detection task. $\mathcal{L}_{total}$ is given as

$$\mathcal{L}_{total} = \lambda_r \times \mathcal{L}_r + \lambda_c \times \mathcal{L}_c \qquad (3)$$

where λ_r and λ_c are the weight coefficients of each task.

The discourse relations of the ten labels explained in Sect. 2.1.2 and chunk labels are identified by softmax. The chunks do not distinguish between hard and soft chunks because the number of hard chunks was smaller than the number of soft chunks in the dataset. Chunk labels are defined as “B” for the start of the chunk, “I” for the inside of the chunk, “E” for the end of the chunk, and “O” for the outside of the chunk.

Word information such as a conjunction is an effective clue to identify discourse relations. The sentence encoder calculates self-attention [14] for a combination of word embeddings of BERT and the embedding of the auxiliary features of the word. The obtained vector and the embedding of the auxiliary features of the sentence are concatenated and given to the bidirectional GRU. The sentence auxiliary features are the sentence and paragraph positions in the document, the sentence position in the paragraph, and the depth in the discourse dependency tree. Since the cause of negative events is often negative and the cause of positive events is often positive [15], emotional polarity information can also effectively determine discourse relations. Hence, the word auxiliary features include sentiment polarity information in addition to part of speech and inflected form.

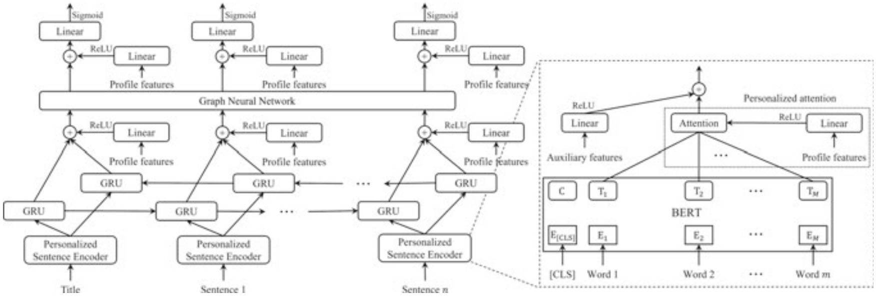


Fig. 4 Interest estimator: BERT_PA_BGRU+_GNN+

3.3 Interest Estimation

Figure 4 overviews the proposed model to estimate the degrees of interest in the topic of a document and each sentence based on the user's profile. The title is inputted before the first sentence because the degree of interest in the title is considered to be the degree of interest in the document's topic. In the personalized sentence encoder, the words of the sentence are given to BERT. Then the personalized attention [16] is calculated for the word embeddings using the profile features as a query. The auxiliary features are the sentence and paragraph positions in the document, the sentence position in the paragraph, and the depth in the discourse dependency tree. Next the sentence vector is given to the bidirectional GRU and the output of the GRU is given to the graph neural network (GNN) [17]. The GNN was introduced based on the analysis of the dataset that the depth of the discourse dependency tree influences the degree of interest. Information is propagated through the dependency structure between sentences with the title as the root node. Profile features are given again before and after the GNN to enhance the effect of the user's profile. Finally, the sentence vectors, which reflect the discourse structure and the user's profile, are given to the output layer with a sigmoid activation function, and the user's interest in each sentence is estimated.

3.4 Interesting Document Selection

The problem of selecting N documents that the user may be interested in from $|D|$ documents with different topics is formulated as an integer linear programming problem as

$$\text{max.} \quad \sum_{k<l \in D} a_k^u a_l^u \left(1 - r_{kl}^d\right) y_{kl}^d \tag{4}$$

Table 1 Variable definitions in the interesting document selection method

x_k^d	Whether document d_k is selected
y_{kl}^d	Whether both d_k and d_l are selected
r_{kl}^d	Similarity between d_k and d_l
N	Maximum number of documents to select
D	Document IDs

s.t.

$$\forall k, l : x_k^d \in \{0, 1\}, \;\; y_{kl}^d \in \{0, 1\}$$

$$\sum_{k \in D} x_k^d \leq N \tag{5}$$

$$\forall k, l : \quad y_{kl}^d - x_k^d \leq 0 \tag{6}$$

$$\forall k, l : \quad y_{kl}^d - x_l^d \leq 0 \tag{7}$$

$$\forall k, l : \quad x_k^d + x_l^d - y_{kl}^d \leq 1 \tag{8}$$

Table 1 explains each variable. a_k^u is the degree of user u's interest in the topic of the document d_k estimated by the interest estimator. r_{kl}^d represents the cosine similarity between the bag-of-words constituting d_k and d_l. Equation 5 is a constraint restricting the number of selecting documents is N or less. Equations 6–8 are constraints that set $y_{kl}^d = 1$ when d_k and d_l are selected.

3.5 Interesting Sentence Extraction

We considered a summarization problem, which extracts sentences that user u may be interested in from the selected N documents and then transmits them by voice within T seconds. The summary must be of interest to the user, coherent, and not redundant. Therefore, we formulated the summarization problem as an integer linear programming problem in which the objective function is defined by the balance between a high degree of interest in the sentences and a low degree of similarity between the sentences with the discourse structure as constraints.

$$\text{max.} \quad \sum_{k \in D_N^u} \sum_{i < j \in S_k} b_{ki}^u b_{kj}^u \left(1 - r_{kij}^s\right) y_{kij}^s \tag{9}$$

s.t.

$$\forall k, i, j : \;\; x_{ki}^s \in \{0, 1\}, \;\; y_{kij}^s \in \{0, 1\}$$

$$\sum_{k \in D_N^u} \sum_{i \in S_k} t_{ki}^s x_{ki}^s \leq T \tag{10}$$

$$\forall k < l : -L \leq \sum_{i \in S_k} x_{ki}^s - \sum_{i \in S_l} x_{li}^s \leq L \tag{11}$$

$$\forall k, i : \quad j = f_k(i), \quad x_{ki}^s \leq x_{kj}^s \tag{12}$$

$$\forall k, m, i \in C_{km} : \quad \sum_{j \in C_{km}} x_{kj}^s = |C_{km}| \times x_{ki}^s \tag{13}$$

$$\forall k, i, j : \quad y_{kij}^s - x_{ki}^s \leq 0 \tag{14}$$

$$\forall k, i, j : \quad y_{kij}^s - x_{kj}^s \leq 0 \tag{15}$$

$$\forall k, i, j : \quad x_{ki}^s + x_{kj}^s - y_{kij}^s \leq 1 \tag{16}$$

Table 2 explains each variable. Here, the i-th sentence of the k-th document is expressed as s_{ki}. b_{ki}^u is calculated based on the degree of user u's interest in the sentence p_{ki}^u which is estimated by the interest estimator and the utterance time t_{ki}^s as $b_{ki}^u = p_{ki}^u \times t_{ki}^s$ to avoid preferential extraction of short sentences. r_{kij}^s represents the cosine similarity between the bag-of-words constituting s_{ki} and s_{kj}. Equation 10 is a constraint restricting the utterance time of the summary to T seconds or less. Equation 11 is a constraint restricting the bias of the number of extracting sentences between documents to L sentences or less. Equation 12 is a constraint in which the parent s_{kj} of s_{ki} in the discourse dependency tree must be extracted when s_{ki} is extracted. Equation 13 is a constraint requiring other sentences in the chunk to be extracted when extracting s_{ki} in a chunk. Equations 14–16 are the constraints that set $y_{kij}^s = 1$ when s_{ki} and s_{kj} are selected.

The maximum bias in the number of extracting sentences between documents L is calculated by the following formulas based on the maximum summary length T, the number of documents N, and the average utterance time of the sentences $\bar{t}$.

Table 2 Variable definitions in the interesting sentence extraction method

x_{ki}^s	Whether sentence s_{ki} is selected
y_{kij}^s	Whether both s_{ki} and s_{kj} are selected
r_{kij}^s	Similarity between s_{ki} and s_{kj}
t_{ki}^s	Utterance time of s_{ki} (seconds)
T	Maximum summary length (seconds)
L	Maximum bias in the number of extracting sentences between documents
$f_k(i)$	Function that returns the parent ID of s_{ki}
D_N^u	IDs of the selected N documents for user u
S_k	Sentence IDs contained in document d_k
C_{km}	Sentence IDs contained in chunk m of d_k

$$L = \left\lfloor \frac{\bar{n}}{\sqrt{N}} + 0.5 \right\rfloor \tag{17}$$

$$\bar{n} = \frac{T}{\bar{t} \times N} \tag{18}$$

$\bar{n}$ represents the expected number of sentences to be extracted from one document. L is the value obtained by dividing $\bar{n}$ by the square root of the number of documents and rounding the result.

4 Experiments

4.1 Discourse Analysis

4.1.1 Experimental Setup

We used the pre-trained BERT model published by the National Institute of Information and Communications Technology[1] (NICT-BERT). This model was trained using BERT$_{\text{BASE}}$ [10] with a vocabulary size of 100,000, which was inputted with text that MeCab[2] [18] morphologically analyzed using the Juman dictionary for all Japanese Wikipedia articles. The dimensions of the GRU hidden layer and linear layer were 128. Adam [19] was used as the optimizer. The evaluation was performed by a tenfold cross-validation, where 9/10 of the articles in each genre were used as training data (1080 articles) and the remaining 1/10 was used as test data (120 articles).

Since the number of discourse relations in the dataset was biased, the classification performance of infrequent discourse relations deteriorated when all the data were used. To suppress the influence of this bias, the sequences of the root-to-leaf sentences of the articles containing at least one target discourse relation were used as a dataset for each discourse relation. *Start* and *Supplement* were excluded as evaluation targets because *Start* is automatically given to sentences whose parents are the root node and *Supplement* is given to those not classified into other discourse relations.

We used articles that contained at least one chunk. The evaluation was performed based on two viewpoints: chunk detection performance and chunk sentence detection performance. The chunk detection performance is F$_1$ [20] when all the chunk ranges match. The chunk sentence detection performance is F$_1$ of the I label when the B and E labels are aggregated into the I label.

[1] https://alaginrc.nict.go.jp/nict-bert/index.html.

[2] https://taku910.github.io/mecab/.

Table 3 Inter-sentence dependency parsing (accuracy)

w/ Sentence position features	0.768
w/o Sentence position features	0.717
Parent is the previous sentence	0.618

Table 4 Chunk detection (F_1)

	Single-task	Multi-task ($\lambda_r = 0.2, \lambda_c = 0.8$)
Chunk	0.605	0.629
Chunk sentence	0.720	0.737

Table 5 Discourse relation classification (F_1)

	Single-task	Multi-task
Result	0.465	0.497 ($\lambda_r = 0.8, \lambda_c = 0.2$)
Cause	0.615	0.640 ($\lambda_r = 0.9, \lambda_c = 0.1$)
Background	0.505	0.510 ($\lambda_r = 0.9, \lambda_c = 0.1$)
Correspondence	0.406	0.417 ($\lambda_r = 0.9, \lambda_c = 0.1$)
Contrast	0.888	0.896 ($\lambda_r = 0.5, \lambda_c = 0.5$)
Topic Change	0.678	0.696 ($\lambda_r = 0.6, \lambda_c = 0.4$)
Example	0.410	0.466 ($\lambda_r = 0.8, \lambda_c = 0.2$)
Conclusion	0.442	0.449 ($\lambda_r = 0.9, \lambda_c = 0.1$)

4.1.2 Experimental Results

Table 3 shows the accuracy of inter-sentence dependency parsing. The baseline shows the performance when the parent is the previous sentence. The model with sentence position features showed an accuracy improvement of at least 5%.

Table 4 shows the chunk detection performance. The performance of the multi-task model was maximum when $\lambda_r = 0.2, \lambda_c = 0.8$ among $(\lambda_r, \lambda_c) \in \{(0.9, 0.1), (0.8, 0.2), \ldots, (0.1, 0.9)\}$. The multi-task model had a higher performance than the single-task model. By also learning the discourse relations, the multi-task model learned that *Contrast* sentences tended to be soft chunks.

Table 5 shows the classification performance of each discourse relation. the evaluation metric was F_1. The results of the multi-task models in the table show the best one among $(\lambda_r, \lambda_c) \in \{(0.9, 0.1), (0.8, 0.2), \ldots, (0.1, 0.9)\}$. The multi-task models exhibited higher performances than the single-task models. Comparing the results of each discourse relation, *Contrast* had the highest performance because sentences with *Contrast* often start with a specific phrase such as “On the other hand.”

Table 6 Interest estimation (accuracy)

	Topic	Sentence
BERT_PA_BGRU+_GNN+	0.701	0.671
BERT_PA_BGRU_GNN	0.688	0.661
BERT_PA_BGRU	0.673	0.649
BERT_BGRU_GNN	0.657	0.630

4.2 Interest Estimation

4.2.1 Experimental Setup

We used data from 1,154 participants who met the following criteria: (1) Answer time of the 6 articles is at least 6 min but less than 20 min. (2) Age is between 20 and 60 years old. (3) Neither occupation type nor industry type is "other." (4) Occupation type is a frequent one.

We used NICT-BERT (explained in Sect. 4.1.1) as the pre-trained BERT model. The dimensions of the GRU hidden layer and the linear layer were 128. Adam was used as the optimizer. ARMAConv [17] of PyTorch Geometric[3] 1.6.3 with the default parameters was used as the GNN. The interest estimator was trained with the labels of the sentences annotated "not interested at all," "not interested," or "not interested if anything" as "0," and the labels of the sentences annotated "very interested," "interested," or "interested if anything" as "1." The evaluation was performed by the tenfold cross-validation where 9/10 of the participants' data was used as the training set, and the remaining 1/10 of the participants' data was used as the test set. Using accuracy as the evaluation metric, we compared the proposed model with the following three models. BERT_PA_BGRU_GNN removed the input of profile features before and after the GNN from BERT_PA_BGRU+_GNN+, BERT_BGRU_GNN removed the personalized attention from BERT_PA_BGRU_GNN, and BERT_PA_BGRU removed the GNN from BERT_PA_BGRU_GNN.

4.2.2 Experimental Results

Table 6 shows the experimental results. The results are divided into "topic" and "sentence." BERT_PA_BGRU+_GNN+ and BERT_PA_BGRU_GNN had higher accuracies than that of BERT_BGRU_GNN. This demonstrates the effectiveness of considering users' profiles. Furthermore, BERT_PA_BGRU_GNN had a higher accuracy than that of BERT_PA_BGRU. This demonstrates the effectiveness of considering the inter-sentence dependencies.

[3] https://pytorch-geometric.readthedocs.io/en/latest/.

4.3 *Personalized Summarization*

4.3.1 Experimental Setup

Using the constructed dataset, we evaluated the performance of the personalized summarization method for dialog scenario planning. We assumed a situation where each of 1,154 users would select three interesting articles from six news articles of different genres. The selected articles were summarized based on their degree of interest, which were transmitted by voice within $T' = 210$ s. Each sentence of the news articles was synthesized by AITalk 4.1[4] to calculate the duration of speech. The maximum summary length T is calculated as $T = T' - (N - 1) \times (q_d - q_s)$, where T' denotes the total utterance time of the primary plan, q_s denotes the pause between sentences, and q_d denotes the pause between documents. Here, $q_s = 1$ second and $q_d = 3$ s. The value obtained by adding q_s to the playback time of the synthesized audio file was set as t^s_{ki}. The integer linear programming problem was solved by the branch-and-cut method[5] [21, 22]. The PULP_CBC_CMD solver of the PuLP[6] 2.4, which is a Python library for linear programming optimization, was used.

The summaries generated by BERT_PA_BGRU+_GNN+, which was trained with the dataset that we constructed in this study, are referred to as interest-based summaries. The summaries generated by BERT_BGRU, which was trained with 100 news articles annotated according to whether each sentence is important or not (Fleiss' kappa of three annotators was 0.546), are referred to as importance-based summaries. Using the data of 1,154 users, we calculated the evaluation metrics described in the next section for each user and compared the average values.

4.3.2 Evaluation Metrics

We propose a metric to evaluate the quality of information transmission. The information transmission quality is calculated based on the efficiency and coherence as

$$\text{QoIT}_{\alpha,\beta,\gamma} = \alpha \times \text{EoIT}_{\beta} + (1 - \alpha) \times \text{CoIT}_{\gamma} \tag{19}$$

Because coherence is considered to be as important as efficiency in the news delivery task, we set $\alpha = 0.5$.

EoIT_{β} is the evaluation metric for efficiency [23]. When C is the coverage of sentences annotated as “very interested,” “interested,” or “interested if anything,” and E is the exclusion rate of the sentences annotated as “not interested at all,” “not interested,” or “not interested if anything,” EoIT_{β} is defined based on the weighted F-measure [20] as

[4] https://www.ai-j.jp/product/voiceplus/manual/.

[5] https://projects.coin-or.org/Cbc.

[6] https://coin-or.github.io/pulp/.

$$\mathrm{EoIT}_{\beta} = \frac{\left(1+\beta^2\right) \times C \times E}{\beta^2 \times C + E} \tag{20}$$

When $\beta = 2$, the exclusion rate is twice as important as the coverage. Compared to textual media, which allows readers to read at their own pace, dialog-based media does not allow users to skip unnecessary information or skim only necessary information while listening. Consequently, we assumed that the exclusion rate is more important than the coverage in information transmission by spoken dialog and set $\beta = 2$.

CoIT_{γ} is the evaluation metric for coherence. When A_d is the accuracy of dependency parsing and F_c is the F-measure of the chunk sentence detection, the CoIT_{γ} is defined as

$$\mathrm{CoIT}_{\gamma} = \gamma \times A_d + (1-\gamma) \times F_c \tag{21}$$

We set $\gamma = 0.8$ because we assumed that correctness of the dependency is more important than the correctness of the chunk.

4.3.3 Experimental Results

Table 7 compares the performance when the actual values of the dataset are given and the performance when the predicted values of the models are given. (1) represents the performance when the sentence interest, topic interest, and discourse structure are all ideal. On the other hand, (5) shows the performance when these are estimated by the proposed models. Comparing (1) and (2) or (1) and (4), revealed at 6–7% difference between the ideal and predicted values of the QoIT when the discourse structure was accurate. On the other hand, when the discourse structure was a prediction, the difference between the ideal and predicted values of the QoIT was almost 20% by comparing (1) and (5). To realize high-quality information transmission, improving the performance of the discourse analysis models should be prioritized. Finally, comparing (2) and (3) verified that the interest-based summaries are more efficient than the importance-based summaries.

5 Conclusion

We proposed a method to generate a personalized summary that may be of interest to each user based on the discourse structure of documents in order to deliver a certain amount of coherent and interesting information for a spoken news dialog system. We constructed a news article corpus with annotations of the discourse structure, users' profiles, and interests in sentences and topics. Our experiments confirmed that the personalized summaries generated by the proposed method transmit information more efficiently than generic summaries generated based on the importance of sentences.

Table 7 Information transmission quality of the summaries ($N = 3$, $T' = 210$)

	Model setings			Evaluation metrics						
	Sentence interest	Topic interest	Discourse	$\mathrm{QoIT}_{0.5,2,0.8}$	EoIT_2	C	E	$\mathrm{CoIT}_{0.8}$	A_d	F_c
(1)	Ideal	Ideal	Ideal	0.855	0.711	0.538	0.806	1.00	1.00	1.00
(2)	Prediction	Ideal	Ideal	0.796	0.592	0.475	0.666	1.00	1.00	1.00
(3)	Importance	Ideal	Ideal	0.779	0.559	0.460	0.628	1.00	1.00	1.00
(4)	Prediction	Prediction	Ideal	0.789	0.579	0.486	0.637	1.00	1.00	1.00
(5)	Prediction	Prediction	Prediction	0.667	0.578	0.489	0.632	0.757	0.767	0.717

In the future work, we plan to devise a method to adaptively generate personalized summaries using the dialog history.

Acknowledgements This work was supported by Japan Science and Technology Agency (JST) Program for Creating STart-ups from Advanced Research and Technology (START), Grant Number JPMJST1912 "Commercialization of Socially-Intelligent Conversational AI Media Service."

References

1. Sappelli M, Chu DM, Cambel B, Graus D, Bressers P (2018) SMART journalism: personalizing, summarizing, and recommending financial economic news. In: The Algorithmic Personalization and News (APEN18) Workshop at ICWSM 18(5):1–3
2. Mani I, Bloedorn E (1998) Machine learning of generic and user-focused summarization. In: Proceedings of the 15th national/10th conference on artificial intelligence/innovative applications of artificial intelligence, pp 820–826
3. Díaz A, Gervás P (2007) User-model based personalized summarization. Inf Process Manage 43(6):1715–1734
4. Yan R, Nie JY, Li X (2011) Summarize what you are interested in: an optimization framework for interactive personalized summarization. In: Proceedings of the 2011 conference on empirical methods in natural language processing, pp 1342–1351
5. Hu P, Ji D, Teng C, Guo Y (2012) Context-enhanced personalized social summarization. In: Proceedings of the 24th international conference on computational linguistics, pp 1223–1238
6. Hirao T, Nishino M, Yoshida Y, Suzuki J, Yasuda N, Nagata M (2015) Summarizing a document by trimming the discourse tree. IEEE/ACM Trans Audio, Speech Lang Process 23(11):2081–2092
7. Kikuchi Y, Hirao T, Takamura H, Okumura M, Nagata M (2014) Single document summarization based on nested tree structure. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, pp 315–320
8. Xu J, Gan Z, Cheng Y, Liu J (2020) Discourse-aware neural extractive text summarization. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 5021–5031
9. Takatsu H, Fukuoka I, Fujie S, Hayashi Y, Kobayashi T (2018) A spoken dialogue system for enabling information behavior of various intention levels. J Jpn Soc Artif Intell 33(1):1–24
10. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 4171–4186
11. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681
12. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing, pp 1724–1734
13. Zhang X, Cheng J, Lapata M (2017) Dependency parsing as head selection. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics, pp 665–676
14. Lin Z, Feng M, dos Santos CN, Yu M, Xiang B, Zhou B, Bengio Y (2017) A structured self-attentive sentence embedding. In: Proceedings of the 5th international conference on learning representations, pp 1–15
15. Oh JH, Torisawa K, Hashimoto C, Kawada T, Saeger SD, Kazama J, Wang Y (2012) Why question answering using sentiment analysis and word classes. In: Proceedings of the 2012

joint conference on empirical methods in natural language processing and computational natural language learning, pp 368–378

16. Wu C, Wu F, An M, Huang J, Huang Y, Xie X (2019) NPA: neural news recommendation with personalized attention. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp 2576–2584
17. Bianchi FM, Grattarola D, Livi L, Alippi C (2021) Graph neural networks with convolutional ARMA filters. IEEE Trans Pattern Anal Mach Intell
18. Kudo T, Yamamoto K, Matsumoto Y (2004) Applying conditional random fields to Japanese morphological analysis. In: Proceedings of the 2004 conference on empirical methods in natural language processing, pp 230–237
19. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: Proceedings of the 3rd international conference for learning representations, pp 1–15
20. Chinchor N (1992) MUC-4 evaluation metrics. In: Proceedings of the 4th conference on message understanding, pp 22–29
21. Mitchell JE (2002) Branch-and-cut algorithms for combinatorial optimization problems. In: Handbook of applied optimization, pp 65–77
22. Padberg M, Rinaldi G (1991) A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems. SIAM Rev 33(1):60–100
23. Takatsu H, Okuda M, Matsuyama Y, Honda H, Fujie S, Kobayashi T (2021) Personalized extractive summarization for a news dialogue system. In: Proceedings of the 8th IEEE spoken language technology workshop, pp 1044–1051