


# Dialog-based multi-item recommendation using automatic evaluation

Euisok Chung  | Hyun Woo Kim | Byunghyun Yoo | Ran Han | Jeongmin Yang | Hwa Jeon Song

Integrated Intelligence Research Section,  
Electronics and Telecommunications  
Research Institute, Daejeon, Republic of  
Korea

## Correspondence

Euisok Chung, Integrated Intelligence  
Research Section, Electronics and  
Telecommunications Research Institute,  
Daejeon, Republic of Korea.  
Email: [eschung@etri.re.kr](mailto:eschung@etri.re.kr)

## Funding information

This research was supported by the  
Electronics and Telecommunications  
Research Institute (ETRI), Grant/Award  
Number: 23ZS1100.

## Abstract

In this paper, we describe a neural network-based application that recommends multiple items using dialog context input and simultaneously outputs a response sentence. Further, we describe a multi-item recommendation by specifying it as a set of clothing recommendations. For this, a multimodal fusion approach that can process both cloth-related text and images is required. We also examine achieving the requirements of downstream models using a pretrained language model. Moreover, we propose a gate-based multimodal fusion and multiprompt learning based on a pretrained language model. Specifically, we propose an automatic evaluation technique to solve the one-to-many mapping problem of multi-item recommendations. A fashion-domain multimodal dataset based on Koreans is constructed and tested. Various experimental environment settings are verified using an automatic evaluation method. The results show that our proposed method can be used to obtain confidence scores for multi-item recommendation results, which is different from traditional accuracy evaluation.

## KEYWORDS

automatic evaluation, multi-item recommendation, multimodal fusion, multiprompt learning

## 1 | INTRODUCTION

Dialog-based product sale systems identify user needs through interactions and recommend a list of suitable products [1, 2]. If dialog processing is handled using a pretrained language model (LM), the multi-item recommendation problem can be approached using the corresponding LM. This requires dealing with multiple complex requirements using a single pretrained LM. In addition, multimodal data processing is required for product items, and an evaluation method is required

because there is no single correct answer to multi-item recommendations. In general, we select a real-world application and examine the core technologies of a dialog-based multi-item recommendation based on the data involved.

Stitch Fix<sup>1</sup> achieved remarkable growth in the fashion industry using a business model that combines large amounts of data and styling expert knowledge [3]. The fashion industry also actively utilizes artificial

<sup>1</sup><https://www.stitchfix.com/>.

intelligence research. DeepFashion [4] was used to conduct a study on clothing recognition by constructing a clothing dataset, and Fashion IQ [5] demonstrated impressive results in a study on clothing image searches using natural language. In this study, our research area included clothing set recommendations through dialog.

Recently, a prompt-based approach that reformulates the input/output of downstream tasks using a large-capacity, pretrained LM [6] has proven to be highly effective among artificial intelligence topics. In addition, studies integrating multimodal information into neural networks have yielded satisfactory results [7, 8]. We used multimodal techniques and a pretrained LM to approach clothing set recommendations. We selected Electra [9] as the pretrained LM. This is because Electra yielded better performance than previously pretrained LMs employing replaced token detection, which can be learned more efficiently than masked LMs.

[10] approached the evaluation problem with multiple answers using one-to-many modeling in natural language generation (NLG). The difficulty in NLG evaluation caused by the one-to-many mapping problem was explored using various methods [11]. Clothing set recommendation topics are not independent of one-to-many mapping problems. This is because there may be numerous recommendable sets of clothes for a dialog context that can be assumed to be the input. In addition, the response prediction for the dialog context should be processed using the same model.

The contributions of this study are as follows.

- We use prompt-based reformulation to approach dialog-based clothing set recommendation and deal with multimodal problems using gate-based early and late fusions.
- Particularly, we present a novel auto-evaluation method to solve the one-to-many mapping problem.
- Finally, we briefly review the multitask learning problem using the dual-weight balancing technique.

## 2 | RELATED WORK

### 2.1 | Prompt-based learning

This technique uses pretrained LM to model the slot-filling probability of texts and performs prediction tasks [6]. Prompts involve two types: cloze prompts [12, 13], which fill in the blanks of a text string, and prefix prompts [14, 15], which successively fill in the string prefix. [16] defined various subprompts and proposed multiprompts for combining them. In this study, prompts were designed based on cloze prompts. In particular, we

approach the multiprompt problem as a multitask learning problem. This is because the composing effects of tasks on learning objectives are implicitly modeled in the self-attention layers.

### 2.2 | Multimodal fusion

Multimodal fusion studies using transformers have attracted increasing attention. [17] proposed a fusion method by concatenating a red-green-blue image and the LiDAR BEV CNN features of an autonomous driving sensor and integrated them using the self-attention of a transformer. [18] presented a multimodal fusion transformer using fusion bottleneck tokens. Studies can be classified based on the layer in which multimodal fusion occurs in neural networks. [19] performed emotion recognition through the fusion of audio, video, and text modality and proposed feature-level early fusion and score-level, late-fusion methods. [8] compared various multimodal fusion models to classify pulmonary embolism cases, reporting the late fusion case to be the best result. Furthermore, early fusion yielded the best performance for Kinetics I3D data in an experiment using multimodal CNN [7]. In this study, we combined multimodal information through a gate that uses the feature value of the input token of a transformer. The gate here acts similarly to the calibration gate in [20]. This was followed by verifying the early and late fusion structures. Early fusion was applied to the word-embedding layer of the pretrained LM, while late fusion was applied before the task head.

### 2.3 | One-to-many mapping problem

In the case of a dialog, there are multiple answers to an input, that is, the context of a conversation, including a user's question. This is known as a one-to-many mapping problem. To solve this problem, [10] modeled a one-to-many relationship using a latent variable. However, an appropriate method to evaluate automatically generated sentences has not yet been proposed. [10] classified the NLG evaluation method into different methods: human-centric evaluation, untrained automatic metrics, and machine-learned metrics. Among these, machine-learned metrics can be used to directly evaluate one-to-many relationships. Sentence-similarity-based methods [21] and regression-based evaluation methods [22] have been proposed. Recently, the similarity between the model output and the correct answer was measured and evaluated using a BERT-based evaluation method [23]. However, the previous studies suffer from a drawback in that

there were cases in which measuring the similarity of the correct answers provided as a reference was difficult. Therefore, using this approach to evaluate the results of a model is challenging, especially if the evaluation area is a clothing set recommendation area rather than an NLG area. In this study, we propose a new automatic evaluation method that can directly measure the relationship between the input and output of a multi-item recommendation model. This method can measure the relationship between multiple items, such as clothing sets, and the corresponding input dialog context.

### 2.4 | Multitask learning

[24] proposed a multitask learning approach for learning multiple objectives with a shared architecture. [25] performed various tasks simultaneously using a unified transformer model (UniT). UniT shares the same model parameters for all tasks and has task-specific output heads. [26] proposed a method to learn representations between multiple NLU tasks through MT-DNN. This method can use the large capacity of cross-task data, thereby achieving a regularization effect through the general representation of new tasks and domains. Multitasking learning also includes the study of the loss of balance between tasks. [27] showed that the normalized random weights for loss values are comparable to the state-of-the-art performance of multitask learning technologies. [28] proposed a hybrid balance approach that trains a model by separating the weights of the feature and loss levels. This study used transformer-based multitask learning for dialog-based multi-item recommendations. We approach

the task of the multi-item recommendation by dividing it into three prompt-based tasks and nine subtasks. In addition, we examine the dual-weight balance in the model and task outputs by referring to the hybrid balance approach of [28].

## 3 | OUR METHODS

This section describes a dialogue-based, multi-item recommendation model using FASCODE. Further, it describes the application of the pretrained LM for response prediction and clothing set recommendations, the feature integration method for images of clothes and clothes information text, and the automatic evaluation of various clothing set recommendations.

### 3.1 | Prompt-based reformulation

We extracted three types of prompt sequences from the FASCODE dialog data for the dialog-based multi-item recommendation system. Figure 1 describes the multi-prompts design. First,  $Task_{US}$  learns the system’s ability to decide, what questions to ask, what answers to provide, and whether to continue recommending the clothing set in the dialog state branch. The input sequence of  $Task_{US}$  consists of a prompt token sequence ( $\langle PS\_US \rangle \langle CS\_US \rangle \langle NS\_US \rangle$ ) for state prediction, a dialog context (INTERACT) consisting of a sequence of word tokens, and an end symbol. The prompt-token type consists of a previous-state token ( $\langle PS\_* \rangle$ ), a current-state token ( $\langle CS\_* \rangle$ ), and a next-state token ( $\langle NS\_* \rangle$ ). The

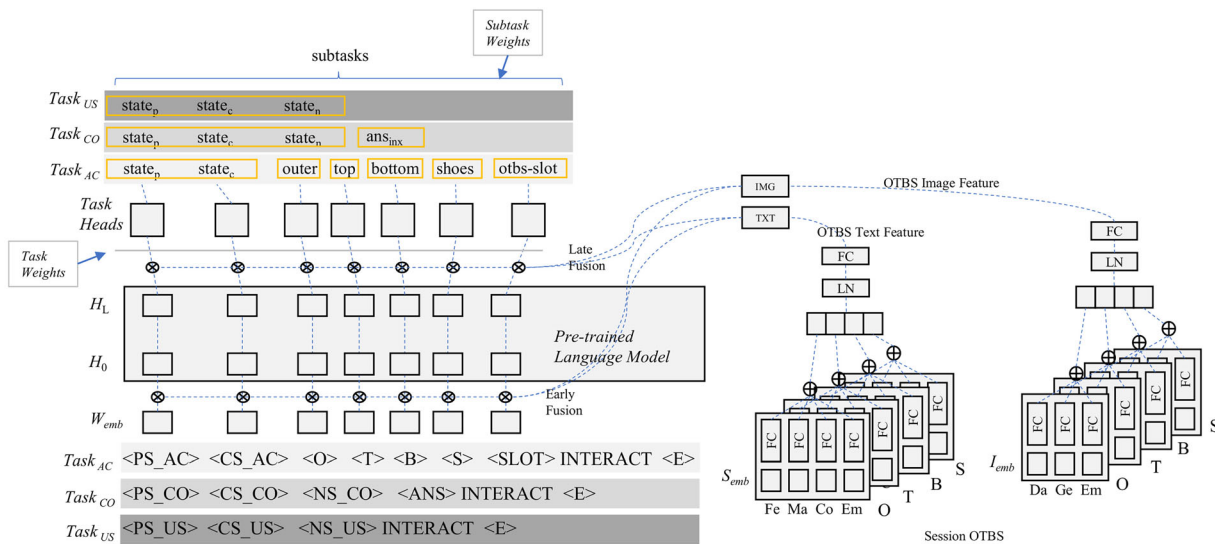


FIGURE 1 Multiprompt-based reformulations. Three prompt-based tasks are used:  $Task_{US}$ ,  $Task_{CO}$ , and  $Task_{AC}$ . These comprise nine subtasks: three states tasks, one answer prediction task, and five cloth recommendation tasks.

outputs, which are the targets to learn, are the previous, current, and subsequent states.

We employed a clustering-based dialog model [29, 30] to predict the responses. Therefore, we assumed that the response sentences were clustered based on the embedding value of the sentence and that each response sentence had a corresponding cluster ID. The cluster ID was used as  $ans_{inx}$ . In the second  $Task_{CO}$ , the prompt token ( $\langle ANS \rangle$ ) was added to the input sequence of  $Task_{US}$ , and the response set ID  $ans_{inx}$  was added as an output.

Clothing recommendation is a process in which the recommendation model predicts the clothing ID for each clothing type using the following tokens: outerwear ( $\langle O \rangle$ ), topwear  $\langle T \rangle$ , bottom wear  $\langle B \rangle$ , and shoe  $\langle S \rangle$ . In the third  $Task_{AC}$ , the prompt token sequence ( $\langle O \rangle \langle T \rangle \langle B \rangle \langle S \rangle$ ) for predicting the clothing set and  $\langle SLOT \rangle$  for predicting the combination of OTBS (four-piece clothes set) are added to the input sequence of  $Task_{US}$ . The next stage is the clothing set recommendation state; therefore, it is excluded from the sequence. The output comprises OTBS clothing set IDs and predicted OTBS combinations.

### 3.2 | Gate-based early and late fusion

In the FASCODE dialog set, the coordinator recommends clothes and receives feedback from users to make changes. In this process, the session OTBS, which consists of approximately four pieces of clothing, is maintained. The multimodal target becomes the Session OTBS, and the right side of Figure 1 describes this process. Text features consist of descriptive categories, such as features, materials, colors, and emotions for one outfit. Each text feature is lexicographed in the state of embedding sentences in advance using the pretrained LM, and all OTBSs are summed after conversion using a fully connected layer (FC). After concatenation, they are converted into text features through layer normalization and FC steps.

The image feature constitutes daily, gender, and embellishment features for each outfit and proceeds similarly to the text feature. These features are obtained by prelearning using deep fashion-learning data and constructed data [31]. Figure 1 shows how IMG and TXT features are combined with  $W_{emb}$  and  $h_L$  in the early and late fusion, respectively. Figure 2 illustrates this process in detail. Herein,  $W_{emb}$  or  $h_L$  acts as a gate for IMG and TXT features using a sigmoidal function, thus demonstrating how they are integrated. The modulation of IMG and TXT based on the sigmoidal function will not proceed if the gate is not applied.

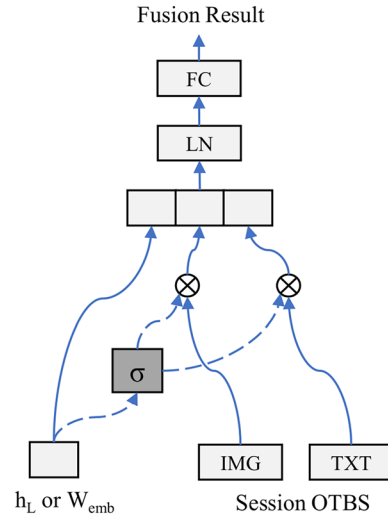


FIGURE 2 Multimodal fusion.

We experimentally verified the early fusion, late fusion, and early-late fusion structures. Early-late fusion is similar to the residual learning approach [32]. The residual method steadily improves performance even when the depth of the neural network increases, and learning and optimization are easy. Early-late fusion enhances the multimodal fusion process by adding fusion steps to the input and output parts of the pre-trained LM. The validity of early-late fusion was confirmed using experimental results.

### 3.3 | Dual-weight balancing

Multitask learning must process multi-item recommendations and response sentence outputs as a single model. We approached this task by dividing it into three prompt-based tasks and nine subsequent subtasks. We used transformer-based multitask learning to integrate various prompt-based task sequences. Feature weight balancing for the three prompt-based tasks was applied to the hidden value-generation part of the shared transformer model. Nine subtasks were applied for weight loss balancing for each loss of prediction of task heads, which were composed of task-specific parameters. Figure 1 shows where weight balancing was applied to the task and sub-task weights.

$$\mathcal{L}_{DWB} = E_{(\lambda, \delta)} [\lambda^\top \mathcal{L}(\mathcal{D}; \delta)]. \quad (1)$$

If the weight loss for the  $m$  subtasks is  $\lambda = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m$ , then the following condition is satisfied:  $\sum_{i=1}^m \lambda_i = 1$ ,  $\lambda_i \geq 0$ . Similarly, if the feature weight for  $n$  prompt-based tasks is  $\delta = (\delta_1, \dots, \delta_n) \in \mathbb{R}^n$ , then the

following condition is required:  $\sum_{i=1}^n \delta_i = 1, \delta_i \geq 0$ . Equation (1) describes the dual-weight balance. This formula is based on [27]. In Equation (1),  $\mathcal{L}(\mathcal{D}; \delta) = (\mathcal{L}_1(\mathcal{D}_1; \delta), \dots, \mathcal{L}_m(\mathcal{D}_m; \delta))$  is a vector describing  $m$  losses for  $m$  subtask data,  $\mathcal{D}_1, \dots, \mathcal{D}_m$ . The loss vector  $\mathcal{L}$  is converted into the loss weight  $\lambda$  through the softmax operation based on considerations of the temperature  $T, \lambda = \text{softmax}(\mathcal{L}(\mathcal{D}; \delta)/T)$ . Herein, a low value of  $T$  results in normal weight loss, whereas a high value of  $T$  results in equal weight loss. When the target of subtask  $i$  is  $y_i$  and the input  $x_i$  is the output of task head TH, loss  $i$  can be described by the cross entropy as follows:  $\mathcal{L}_i(\mathcal{D}_i; \delta) = \text{CE}(y_i, x_i = \text{TH}(\mathcal{D}_i; h_\delta^i))$ . If the data  $\mathcal{D}_i$  of subtask  $i$  are classified as a prompt-based task  $t$ , the data are described as  $\mathcal{D}_i^t$ . The input of the task head TH is the output  $h_\delta^t$  of the pretrained LM (PLM) converting the task data  $\mathcal{D}_i^t$ , which is described as  $h_\delta^t = \delta_i \cdot \text{PLM}(\mathcal{D}_i^t)$ . Herein, the feature weight  $\delta_i$  was applied as the weight of task  $t$ .

Loss weight balancing is determined dynamically when softmax uses a low  $T$  and is forcibly processed as an equal loss weight when using a high  $T$ . Feature weight balancing is treated as a hyperparameter, thus allowing weights to be assigned directly to each prompt-based task. This can be verified by adding a random-weight policy to each balancing step in the experiments.

### 3.4 | Automatic evaluation

In the case of a dialog, there are multiple answers to the input. This is known as a one-to-many mapping problem. In this study, we propose a new automatic evaluation method that can directly define the relationship between the input and output of a multi-item recommendation model. This method can measure the relationship between multiple items, such as clothing sets, and the corresponding input dialog context.

This study uses a model that outputs correlation values for the OTBS prediction and the interaction context of [33]. Figure 3 shows the recommender receiving the interaction context as input and predicting the OTBS. To evaluate this, the input and output sequences were first converted into multiple input sequences based on the rater's input format. The corresponding input sequences have a logit value through the rater, and the sum of all logit values becomes the evaluation value of the recommendation result.

The right-hand side of Figure 3 shows a model (rater) that passes through an independent learning process. First, we created training data from the FASCODE data and divided them into two sequence types:  $\langle \text{CLS} \rangle \text{CONTEXT} \langle \text{SEP} \rangle \text{ITEM\_TEXT} \langle \text{SEP} \rangle$  and  $\langle \text{CLS} \rangle \text{ITEM\_TEXT1} \langle \text{SEP} \rangle \text{ITEM\_TEXT2} \langle \text{SEP} \rangle$ . The former describes the relationship between the conversation context and the clothes item, whereas the latter describes the relationship between the clothes pairs constituting the clothing set. Herein,  $\langle \text{CLS} \rangle$  is a special symbol for obtaining the logit value of the input sequence, and  $\langle \text{SEP} \rangle$  is a special symbol for separating the input type. Both inputs were integrated with each image feature to perform multimodal processing. NONE + ITEM\_IMG\_FEAT and ITEM\_IMG\_FEAT1 + ITEM\_IMG\_FEAT2 were processed using the late fusion method. Learning proceeds such that positive and negative data can be separated using a binary classifier.

The raters' learning results were verified using FASCODE-EVAL. The evaluation set quantified the system performance through the correlation between the conversation context, relevance score of the clothing set, and logit score output by the system. The left-hand side of Figure 4 shows the correlation between relatedness, which is the average score of the evaluators, and the logits of the system. This correlation graph yielded a Spearman's  $\rho$  equal to 0.4. This study assumed that

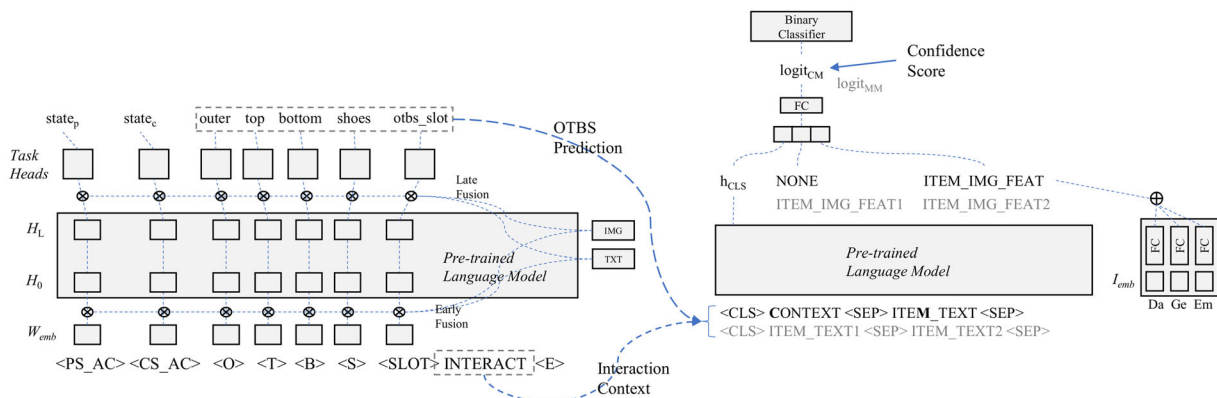


FIGURE 3 Automatic evaluation: recommender and rater.

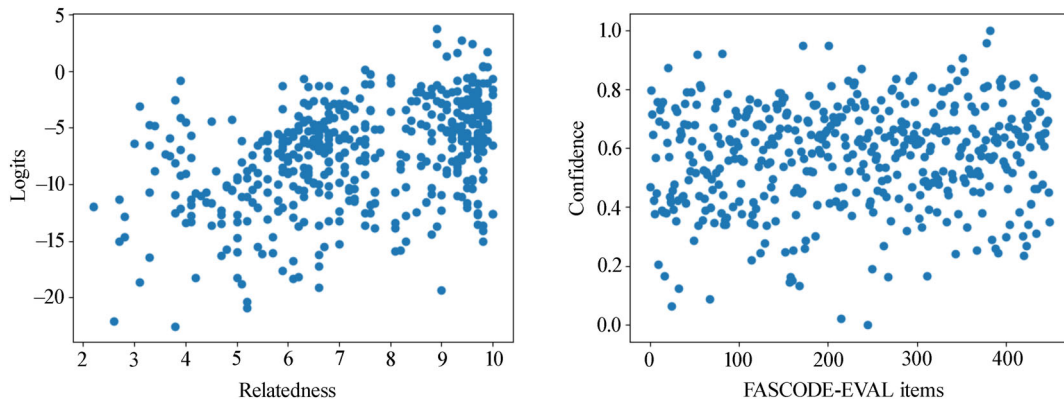


FIGURE 4 Correlation evaluation of rater using FASCODE-EVAL and scaled confidence value.

a rater at this level is similar to the evaluation level of people when evaluating the input/output of the system.

The rater's output was adjusted to a logit value (in the range of 0–1.0). This can be performed using the minimum ( $\min_v$ ) and maximum logit values ( $\max_v$ ) of FASCODE-EVAL. Subsequently, if  $\text{logits}_{\text{avg}}$  is the average of the rater results for the evaluation set, the scaled confidence value can be calculated as  $(\text{logits}_{\text{avg}} - \min_v) / (\max_v - \min_v)$ . The right side of Figure 4 shows the 450 confidence value pairs. The evaluation set was designed evenly from zero to 10, and the graph was effective.

## 4 | EXPERIMENTS

### 4.1 | Dataset

The fashion coordination dataset (FASCODE/Fashion CODE)<sup>2</sup> is composed of FASCODE data, such as the clothing set recommendation dialog set, images, and descriptions of clothes, and FACODE-EVAL, which can evaluate the relationship between the recommended context and clothing set. This dataset has been used in interactive recommendation systems [34], multimodal clothing recommendations [33], and context-based subword embedding technology [30].

The dialog set of FASCODE data comprises 7236 dialog sets with an average of eight to nine turns and 2599 clothing items. The dialog comprises the process associated with the determination of the clothing set between the user and the system. The system recommends clothing sets through interactions with the user, and the user completes the desired clothing set through various types of feedback such as positive, negative, and change requests. The dataset was built by several participants,

TABLE 1 An example of the FASCODE data dialog set.

Task	Utterance	State
<CO>	Hello, how can I help you?	CO_INTRO
<US>	My brother is getting married.	USER_UT
<US>	Please show me a calm and neat outfit.	USER_UT
<AC>	JP-076 BL-027 PT-027 SE-004	AC_OTBS
<CO>	It is a calm and luxurious blouse and trouser coordination.	CO_EXP
<US>	Change your outerwear to a jacket.	USER_FAIL
:	:	:
<CO>	I am really glad you liked it.	CO_SUCCESS
<CO>	Thank you for using it.	CO_CLOSING

Note: The utterance is the result of translation from Korean to English.

who assumed the roles of the system and user. Furthermore, 100 user profiles and 329 time, place, and occasion (TPOs) were used to construct the data. Table 1 is an example of a dialog set, which consists of the system utterance (<CO>), user utterance (<US>), and clothing recommendations (<AC>). The dialog response function can learn <CO> and <AC> as a target and target for clothes recommendation, respectively. Clothing items with IDs are expressed in terms of text, which comprises shape, material, color, and emotion and includes images of clothes.

FASCODE-EVAL used in this study consisted of multiple triples. A triple is expressed in the form of a conversation context, clothes set, and evaluation score using the average value of a pair consisting of the dialog context and clothes set. FASCODE-EVAL consisted of 450 triples. It was constructed in the form of an absolute evaluation of the relationship between a dialog context and clothes set with the help of 10 evaluators on a scale of 0–10. In addition, it was influenced by WordSim353 [35], which is

<sup>2</sup><https://fashion-how.org/ETRI/board.html>.

an evaluation set that measures the similarity between words.

Electra [9] was used as the pretrained LM. The small model of KoElectra [36] was used for the dialog-based multi-item recommendation model because Korean FAS-CODE data were used for the experiments in this study.

## 4.2 | Settings

### 4.2.1 | Accuracy evaluation

To train the recommender, 6211 and 1025 dialog sets were used as the training and evaluation sets, respectively. Each utterance has various hierarchically structured functional tags, which were reduced to 16 states in this study. The OTBS slot was determined as follows.

- state = CO\_ASK, CO\_FAIL, CO\_EXP, CO\_CLOSING, CO\_INTRO, CO\_HELP,..., USER\_SUCCESS, USER\_UT, USER\_FAIL, and AC\_OTBS
- slot = O, T, B, S, OT, OB, OS, TB, TS, BS, OTB, OTS, OBS, TBS, and OTBS

This state was used as a label for the state sequence of a task. The corresponding subtasks were US\_S, CO\_S, and AC\_S. The slot was used as a label for the AC\_SLOT subtask and simultaneously had a masking role for the OTBS clothing set of the recommender.

The recommender included a subtask (CO\_A) that predicted the answer in a conversation. We extracted 29095 response sentences from the FASCODE data and classified them into 1000 sentence sets using k-means clustering.<sup>3</sup> Therefore, each answer sentence comprised a class ID, and CO\_A evaluated the prediction accuracy for this class ID.

Furthermore, OTBS prediction proceeded with subtasks *O*, *T*, *B*, and *S*. We used 1160 outers, 671 tops, 639 bottoms, and 129 shoes as prediction labels. As described in the previous section, OTBS prediction has more than one correct answer. Therefore, the accuracy evaluation of OTBS suffers from a drawback. If the system recommends similar clothes that are different from the correct answer, it is treated as an error.

### 4.2.2 | Confidence evaluation

The raters subsequently study the entire corpus. This is because sufficient data are required to learn the relationship between the dialog context and the clothing set. We

used the rater to evaluate the various methods of the recommender using the confidence policy. Section 3.4 describes the raters' detailed settings.

### 4.2.3 | Parameters

A weight decay  $1e-4$ , learning rate  $2e-5$ , and Adam epsilon value of  $1e-8$  were used for the learning setup. A batch size of 16 was used in the training, and the training and evaluation were reviewed at time steps of 100,000.

### 4.2.4 | Comparison

The multimodal approach experiment consisted of early (E), late (L), and early-late (E\_L) fusion, in which both fusion methods were applied simultaneously. In the gate-based multimodal fusion, the experiment was compared with G based on whether the gate was applied. The basic setting of the experiment was applied as a temperature-based (T)-based softmax according to the loss value caused by the estimation error of the model. A weight loss was applied to T5, whereas equal weights were applied to T100. In the case of feature weights, the experiment was described as US (U), CO (C), and AC tasks (A). In the case of U33\_C33\_A33, the same weight of 0.33 was applied to all cases, and the weights were 0.3, 0.1, and 0.6 for U3\_C1\_A6.

In this study, tests for  $\lambda$  and  $\delta$  in Equation (1) and the objective function of the multiprompts model were briefly conducted through experiments using equal, random, and loss-based weights.

## 4.3 | Results

### 4.3.1 | Accuracy evaluation

Figure 5 shows that the late-fusion (L\_G\_T5\_\*) experiment produced satisfactory results in the accuracy evaluation of subtasks for CO\_A. This is different from the case in which the OTBS early and late (E\_L\_G\_T5\_\*) experimental results were the best. The clothing was predicted directly in the OTBS task. The experimental results demonstrated that performance was improved by integrating early and late fusion. In the state sequence experiments US\_S, CO\_S, and AC\_S, the accuracy improved in the learning stage. US\_S and CO\_S are accuracy evaluations for three consecutive states, including the prediction of the next input context state. The experimental results for state sequence prediction yielded an average accuracy of >80%. This means that the level of prediction

<sup>3</sup><https://github.com/DwangoMediaVillage/pqkmeans>.

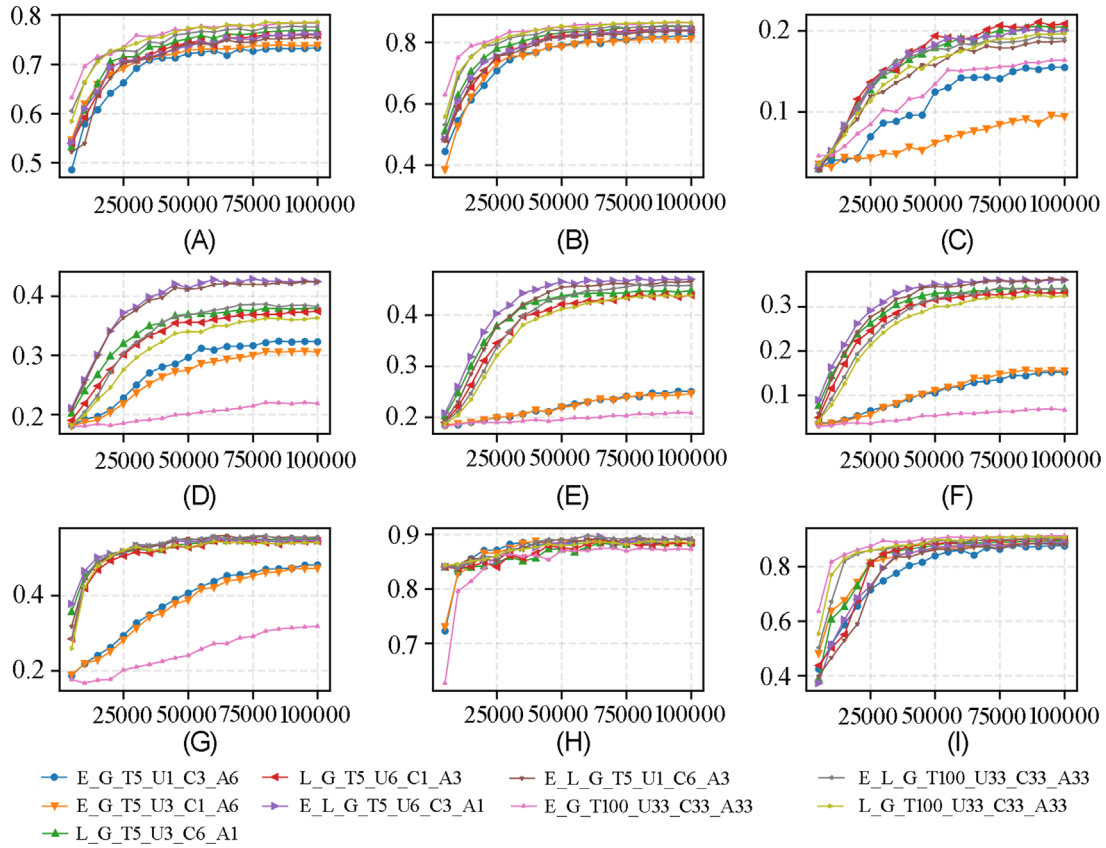


FIGURE 5 Accuracy evaluation: (A) US\_S, (B) CO\_S, (C) CO\_A, (D) O, (E) T, (F) B, (G) S, (H) AC\_SLOT, and (I) AC\_S.

of the next dialog state for the current input in the clothing set recommendation dialog system was  $>80\%$ . Overall, the early fusion ( $E_*$ ) experiments performed poorly. This implies that it is inappropriate to integrate multimodal information before the self-attention layer. When a significance analysis was performed for each model on the accuracy results according to the learning stage, statistically significant experimental results were confirmed in all experiments, except for the AC\_SLOT task.

#### 4.3.2 | Confidence evaluation

In the auto-evaluation experiment, most of the training was performed similarly, except for the  $E_G T100_*$  and  $L_G T100_*$  experiments. However, Figure 6 shows that the early fusion weight loss ( $E_G T5_*$ ) experiment learned quickly. However, because this is a different result from the accuracy evaluation, a rapid performance improvement may not yield a good model. In the confidence evaluation,  $E_L G T100$  performed poorly in the early learning stage but gradually improved and produced the best performance of 0.805. The accuracy experiment in Figure 5 shows that  $E_L G T100$  is among the top three most-featured models (Appendix A) for each task.

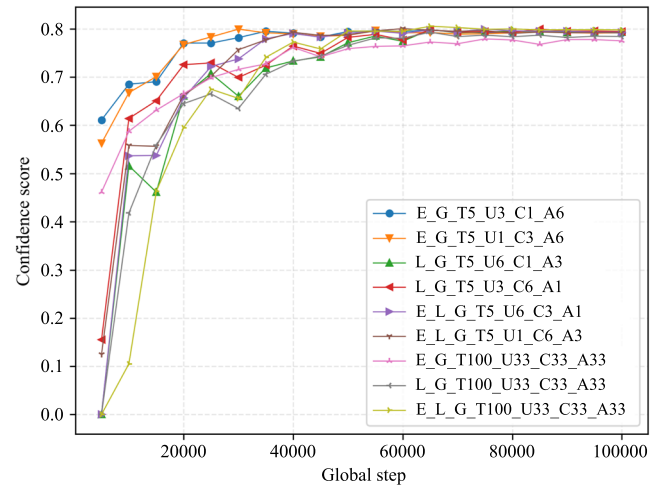


FIGURE 6 Confidence evaluation.

Therefore, we conclude that the results of the confidence evaluation are reasonable and that  $E_L G T100$  is the best model. [7] reported that the performance of early fusion was better than that of late fusion. However, these two fusion methods have not yet been tested simultaneously. Therefore, the results of this study suggest that early late fusion is a valid method for multimodal integration.



### 4.3.3 | Task ablation experiments

The training data for the recommender consisted of three prompt-based tasks. Figure 1 illustrates the US, CO, and AC tasks. We conducted an ablation test on the CO\_A subtasks in CO and a confidence evaluation of the AC. The experiment was conducted using equal weights of  $E\_L\_*$ , and Table 2 lists the results. In the confidence test, performance deteriorated when the tasks were excluded. However, for the CO\_A answer class prediction experiment, the performance improved when the tasks were excluded. This means that other tasks interfered with the learning process for CO\_A. Task ablation experiments on CO\_A show that independent learning for CO\_A and separation of the corresponding model are crucial. In addition, the t-test results showed a significant performance improvement when all tasks were integrated.

### 4.3.4 | Dual-weight balancing

Table 3 presents the dual-weight balancing experiments for random weight (rw), lost weight (lw), and equal weight (ew). The dual-weight balancing tests were

TABLE 2 Task ablation experiments.

Tasks	Accuracy (CO_A)	p-value
CO + US + AC	0.193	-
CO + US	0.203	0.0001
CO	0.205	0.4775
Tasks	Confidence	p-value
AC + US + CO	0.805	-
AC + US	0.796	1.6e-6
AC	0.795	0.8979

Note: The p-value was obtained using a paired t-test between the current row and the previous row test.

TABLE 3 Dual-weight balancing, feature weight balancing (FWB), and loss weight balancing (LWB).

FWB	LWB	Acc. (CO_A)	Conf.	p-value
rw	lw	0.196	0.784	-
ew	lw	0.195	0.786	0.1663
rw	ew	0.194	0.792	0.0212
rw	rw	0.194	0.794	0.4007
ew	ew	0.194	0.802	3.0e-5
ew	rw	0.192	0.804	0.0566

Note: A t-test was conducted with the upper row only for the confidence results.

conducted using  $E\_L\_G\_*$ . lw applied T5 to dynamically determine the weight  $\lambda$  of (1) according to the loss values, and ew applied T100 to all loss values or tasks. rw randomly determined the weight  $\lambda$  or  $\delta$  in (1).  $FWB_{rw} + LWB_{lw}$  exhibited the best performance in CO\_A experiments. However, these results were similar to those of the other experiments, and the t-test did not yield statistically significant results. Similar to the conclusion that the multitask learning approach, which is the result of task ablation, is inappropriate, the CO\_A experiment was not related to dual-weight balancing. In the confidence evaluation,  $FWB_{ew} + LWB_{rw}$  exhibited the best performance. The results of this experiment are similar to those in [27]. The authors of this study argued that random weights (rw) are a competitive technique in multitask learning and the experimental results of the present study support this. However, because of the complex and diverse task composition of this study, the equal-weight technique yielded a performance similar to that of the random-weight approach. The t-test results yielded a significant performance improvement over  $FWB_{ew} + LWB_{ew}$ .

### 4.3.5 | Gate ablation test

We conducted an experiment in which the gates were removed using gate-based multimodal fusion. Table 4 lists the best settings and evaluation results for all experiments, including the gate removal setting. The experiment using the gate yielded better performance in all experiments, except for  $E\_L\_T100$  in S and  $E\_L\_T5$  in the AC\_SLOT experiment. The common feature between S and AC\_SLOT was that their class sizes were smaller than those in the other experiments. Thus, it can be concluded that the gate approach is more effective when the prediction class is large and complex.

TABLE 4 Gate ablation in multimodal fusion.

Tasks	Best settings	Results
US_S	$L\_G\_T100$	0.785
CO_S	$L\_G\_T100$	0.866
CO_A	$L\_G\_T5, L\_T5$	0.210
O	$E\_L\_G\_T5$	0.429
T	$E\_L\_G\_T5$	0.471
B	$E\_L\_G\_T5$	0.363
S	$E\_L\_T100$	0.563
AC_SLOT	$E\_L\_T5$	0.904
AC_S	$E\_G\_T100$	0.913
CONF	$E\_L\_G\_T100$	0.805

TABLE 5 Comparison of execution times of models.

Tasks	Time (min:s)
L_G_T5_U6_C1_A3	11:41
E_G_T100_U33_C33_A33	11:51
L_G_T5_U3_C6_A1	11:53
E_G_T5_U1_C3_A6	11:58
E_G_T5_U3_C1_A6	12:01
L_G_T100_U33_C33_A33	12:01
E_L_G_T100_U33_C33_A33	12:35
E_L_G_T5_U1_C6_A3	12:39
E_L_G_T5_U6_C3_A1	12:40

#### 4.3.6 | Time complexity

Table 5 shows the results of sorting the average execution time for each model in the test set for the accuracy and confidence evaluations. Experimental results show that the early-late fusion model requires a slightly longer execution time than the single fusion model when separated into two groups according to the execution time.

## 4.4 | Discussion

To evaluate multiple-item recommendations, [37] suggested novelty and diversity in addition to accuracy evaluation. However, this has not yet been presented as a single integrated evaluation standard. In addition, novelty and diversity are inappropriate evaluation criteria because of the item-type limitations of this study. We proposed an automatic evaluation method for multi-item recommendations used in confidence evaluation. For the accuracy-based evaluation, Table 4 shows that the evaluation value of the recommendation system can be obtained by multiplying only the best results of the O, T, B, and S tasks. The value was 0.0411, which is extremely low. Another problem is that for accuracy-based evaluations, similar item recommendations or various results that users agree with are considered incorrect answers. However, for confidence evaluations, the automatic evaluation of multiple items was enabled by the learnable rater, and its validity was verified based on comparisons with the accuracy-based evaluation results.

## 5 | CONCLUSIONS AND FUTURE WORK

In this study, we examined dialog-based clothing set recommendations for multiprompt learning using a

pretrained LM. We presented a multimodal fusion approach by switching to multitask learning. Specifically, we present an auto-evaluation approach that uses the rater model as an alternative to the one-to-many mapping problem. Using the evaluation method, we proposed an approach to obtain the confidence score for the multi-item recommendation result, which is distinct from the traditional accuracy evaluation. In the future, we will introduce a more in-depth multitasking learning technique to solve the problem of performance differences between tasks based on the experimental environment.

### ACKNOWLEDGEMENTS

This work was supported by an Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean Government (23ZS1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence Systems).

### CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

### ORCID

Euisok Chung  <https://orcid.org/0000-0001-5091-2508>

### REFERENCES

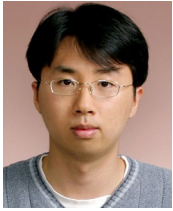
1. D. Pramod and P. Bafna, *Conversational recommender systems techniques, tools, acceptance, and adoption: a state of the art review*, *Expert Syst. Appl.* **203** (2022), 117539.
2. J. Konstan and L. Terveen, *Human-centered recommender systems: origins, advances, challenges, and opportunities*, *AI Mag.* **42** (2021), no. 3, 31–42.
3. K. Zielnicki, *Simulacra and selection: clothing set recommendation at stitch fix*, (Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France), 2019, pp. 1379–1380.
4. Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, *Deep Fashion: Powering robust clothes recognition and retrieval with rich annotations*, (IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA), 2016, DOI [10.1109/CVPR.2016.124](https://doi.org/10.1109/CVPR.2016.124).
5. H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, *Fashion IQ: A new dataset towards retrieving images by natural language feedback*, (IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA), 2021, pp. 11307–11317.
6. P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, *Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing*, *arXiv preprint*, 2021, DOI [10.48550/arXiv.2107.13586](https://doi.org/10.48550/arXiv.2107.13586).
7. K. Gadzicki, R. Khamsehashari, and C. Zetsche, *Early vs late fusion in multimodal convolutional neural networks*, (IEEE 23rd International Conference on Information Fusion, Rustenburg, South Africa), 2020, pp. 1–6.
8. S. Huang, A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren, *Multimodal fusion with deep neural networks for*

- leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection, *Sci. Reports* **10** (2020), 22147.
9. K. Clark, T. Luong, Q. V. Le, and C. Manning, *ELECTRA: Pre-training text encoders as discriminators rather than generators*, (8th International Conference on Learning Representations, Virtual Conference), 2020.
  10. S. Bao, H. He, F. Wang, H. Wu, and H. Wang, *PLATO: pre-trained dialogue generation model with discrete latent variable*, (Proc. 58th Annual Meeting of the Association for Computational Linguistics), 2020, pp. 85–96.
  11. A. Celikyilmaz, E. Clark, and J. Gao, *Evaluation of text generation: a survey*, arXiv preprint, 2020, DOI [10.48550/arXiv.2006.14799](https://doi.org/10.48550/arXiv.2006.14799).
  12. F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, *Language models as knowledge bases?* (Proceedings of conference on EMNLP-IJCNLP, Hong Kong, China), 2019, pp. 2463–2473.
  13. L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang, *Template-based named entity recognition using BART*, arXiv preprint, 2021, DOI [10.48550/arXiv.2106.01760](https://doi.org/10.48550/arXiv.2106.01760)
  14. X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, *A unified MRC framework for named entity recognition*, (Proc. 58th Annual Meeting of the Association for Computational Linguistics, Online), 2020, pp. 5849–5859.
  15. B. Lester, R. Al-Rfou, and N. Constant, *The power of scale for parameter-efficient prompt tuning*, arXiv preprint, 2021, DOI [10.48550/arXiv.2104.08691](https://doi.org/10.48550/arXiv.2104.08691).
  16. X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, *PTR: prompt tuning with rules for text classification*, arXiv preprint, 2021, DOI [10.48550/arXiv.2105.11259](https://doi.org/10.48550/arXiv.2105.11259).
  17. A. Prakash, K. Chitta, and A. Geiger, *Multi-modal fusion transformer for end-to-end autonomous driving*, arXiv preprint, 2021, DOI [10.48550/arXiv.2104.09224](https://doi.org/10.48550/arXiv.2104.09224).
  18. A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, Attention bottlenecks for multimodal fusion, *Proc. NIPS*, **34** (2021), 14200–14213.
  19. J. D. S. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, *Multimodal fusion with deep neural networks for audio-video emotion recognition*, arXiv preprint 2019, DOI [10.48550/arXiv.1907.03196](https://doi.org/10.48550/arXiv.1907.03196).
  20. Y. Lu, J. Zeng, J. Zhang, S. Wu, and M. Li, *Attention calibration for transformer in neural machine translation*, (Proceedings of ACL-IJCNLP, Online), 2021, pp. 1288–1298.
  21. R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, *Skip-thought vectors*, arXiv preprint, 2015, DOI [10.48550/arXiv.1506.06726](https://doi.org/10.48550/arXiv.1506.06726).
  22. L. Logeswaran and H. Lee, *An efficient framework for learning sentence representations*, (Proceedings of International Conference on Learning Representations, Vancouver, Canada), 2018.
  23. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, *BERTScore: evaluating text generation with BERT*, (Proceedings of International Conference on Learning Representations, Online), 2020.
  24. F. Heuer, S. Mantowsky, S. S. Bukhari, and G. Schneider, *MultiTask-Centernet (MCN): Efficient and diverse multitask learning using an anchor free approach*, (IEEE/CVF International Conference on Computer Vision Workshops, Montreal, Canada), 2021, pp. 997–1005.
  25. R. Hu and A. Singh, *UniT: Multimodal multitask learning with a unified transformer*, (IEEE/CVF International Conference on Computer Vision, Montreal, Canada), 2021, pp. 1439–1449.
  26. X. Liu, P. He, W. Chen, and J. Gao, *Multi-task deep neural networks for natural language understanding*, (Proceedings of ACL, Florence, Italy), 2019, pp. 4487–4496.
  27. B. Lin, F. Ye, Y. Zhang, and I. W. Tsang, Reasonable effectiveness of random weighting: A litmus test for multi-task learning, arXiv preprint, 2021, DOI [10.48550/arXiv.2111.10603](https://doi.org/10.48550/arXiv.2111.10603)
  28. L. Liu, Y. Li, Z. Kuang, J. Xue, Y. Chen, W. Yang, Q. Liao, and W. Zhang, *Towards impartial multi-task learning*, (Proceedings of International Conference on Learning Representations), 2021.
  29. R. C. Gunasekara, D. Nahamoo, L. C. Polymenakos, D. E. Ciaurri, J. Ganhotra, and K. P. Fadnis, *Quantized dialog—a general approach for conversational systems*, *Comput Speech Lang.* **54** (2019), 17–30.
  30. E. Chung, H. W. Kim, and H. J. Song, *Sentence model based subword embeddings for a dialog system*, *ETRI J.* **44** (2022), 599–612.
  31. M. Park, H. J. Song, and D. Kang, *Imbalanced classification via feature dictionary-based minority oversampling*, *IEEE Access* **10** (2022), 34236–34245.
  32. K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, (IEEE Conference on Computer Vision and Pattern Recognition), 2016, pp. 770–778.
  33. E. Chung, H. W. Kim, M. Park, and H. J. Song, *Multi-modal approach for FASCODE-EVAL*, (Annual Conference on Human and Language Technology), 2021, pp. 514–517.
  34. E. Chung, H. W. Kim, H. Oh, and H. J. Song, *Dataset for interactive recommendation system*, (Annual Conference on Human and Language Technology), 2020, pp. 481–485.
  35. E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, *A study on similarity and relatedness using distributional and WordNet-based approaches*, (Proceedings of NAACL, Boulder, CO, USA), 2009, pp. 19–27.
  36. J. Park, Koelectra: Pretrained electra model for Korean, 2020. <https://github.com/monologg/KoELECTRA>
  37. A. Jain, P. K. Singh, and J. Dhar, *Multi-objective item evaluation for diverse as well as novel item recommendations*, *Expert Syst. Appl.* **139** (2020), 112857.

## AUTHOR BIOGRAPHIES



**Euisok Chung** received his BS degree in 1997 in Computer Science from Soongsil University, Seoul, Republic of Korea, and his MS degree in 1999 in Computer Science from Yonsei University, Seoul. Since 1999, he has been working with ETRI in Daejeon, Republic of Korea. His current research interests include natural language processing, machine learning, and spoken dialogue systems.



**Hyun Woo Kim** received his BS and MS degrees in Electrical Engineering from Seoul National University (SNU), Seoul, Republic of Korea, in 2001 and 2003, respectively. Since 2003, he has been with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea, where he is currently a Principal Member of the Engineering Staff. His research interests include speech signal processing, meta-learning, and machine learning.

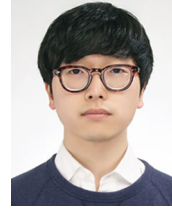


**Byunghyun Yoo** received his PhD degree in Mechanical Engineering from the Korea Advanced Institute of Science and Technology (KAIST) in 2019. Since 2019, he has been with the Electronics and Telecommunications Research Institute (ETRI), Republic of Korea, where he is currently a Senior Researcher with the Integrated Intelligence Research Section and the Intelligence Information Research Division. His research interests include multi-agent reinforcement learning (MARL) and artificial general intelligence (AGI).



**Ran Han** received the BS, MS, and PhD degrees in Electrical and Electronics Engineering from Yonsei University, Seoul, Republic of Korea, in 2008, 2010, and 2015, respectively. She was a Senior Engineer at Samsung Electronics, Suwon, Republic of Korea from 2015 to 2017. In 2017, she joined the Electronics and Telecommunications Research Institute (ETRI) in Daejeon, Republic of Korea, where she is currently a Senior Researcher of the Artificial

Intelligence Group. Her research interests include digital signal processing, pattern recognition, and machine learning.



**Jeongmin Yang** received his BS and MS degrees in Electrical Engineering from the School of Electrical Engineering of KAIST, Daejeon, Republic of Korea in 2012 and 2014, respectively. Since 2014, he has worked at the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea, where he is currently a senior researcher. His primary research interests include meta-learning and multi-agent reinforcement learning.



**Hwa Jeon Song** received the BS, MS, and PhD degrees in Electronics Engineering from Pusan National University, Republic of Korea in 1993, 1995, and 2005, respectively. From 1995 to 2001, he was a Researcher with Hyundai Motor Company. Since 2010, he has been a Principal Researcher with the Electronics and Telecommunications Research Institute (ETRI). His research interests include speech recognition, multimodal representations, and artificial general intelligence (AGI).

**How to cite this article:** E. Chung, H. W. Kim, B. Yoo, R. Han, J. Yang, and H. J. Song, *Dialog-based multi-item recommendation using automatic evaluation*, ETRI Journal (2023), 1–13. DOI [10.4218/etrij.2022-0333](https://doi.org/10.4218/etrij.2022-0333).

## APPENDIX A: TOP 3 IN ACCURACY EVALUATION

Table A1 shows the top three accuracy test results. Herein, the E\_L\_G\_T100\_U33\_C33\_A33 model was included in the top three results with the highest frequency eight times.

TABLE A1 Top 3 accuracy evaluation outcomes.

Tasks	Top 3
US_S	(L_G_T100_U33_C33_A33, 0.7855), (E_G_T100_U33_C33_A33, 0.7832), (E_L_G_T100_U33_C33_A33, 0.7771)
CO_S	(L_G_T100_U33_C33_A33, 0.8665), (E_G_T100_U33_C33_A33, 0.8649), (E_L_G_T100_U33_C33_A33, 0.8546)
CO_A	(L_G_T5_U6_C1_A3, 0.2108), (L_G_T5_U3_C6_A1, 0.2061), (E_L_G_T5_U6_C3_A1, 0.201)
O	(E_L_G_T5_U6_C3_A1, 0.4294), (E_L_G_T5_U1_C6_A3, 0.4247), (E_L_G_T100_U33_C33_A33, 0.3868)
T	(E_L_G_T5_U6_C3_A1, 0.471), (E_L_G_T5_U1_C6_A3, 0.4659), (E_L_G_T100_U33_C33_A33, 0.4599)
B	(E_L_G_T5_U1_C6_A3, 0.3634), (E_L_G_T5_U6_C3_A1, 0.3609), (E_L_G_T100_U33_C33_A33, 0.3438)
S	(E_L_G_T5_U1_C6_A3, 0.5594), (E_L_G_T100_U33_C33_A33, 0.5593), (E_L_G_T5_U6_C3_A1, 0.5519)
AC_SLOT	(E_L_G_T100_U33_C33_A33, 0.8977), (E_L_G_T5_U6_C3_A1, 0.8958), (E_L_G_T5_U1_C6_A3, 0.8925)
AC_S	(E_G_T100_U33_C33_A33, 0.9136), (L_G_T100_U33_C33_A33, 0.9102), (E_L_G_T100_U33_C33_A33, 0.9042)