Interactive User Interface for Dialogue Summarization

Jeesu Jung*
jisu.jung5@gmail.com
Chungnam National University
Yuseong-gu, Daejeon, Republic of Korea

Hyein Seo*
hyenee97@gmail.com
Chungnam National University
Yuseong-gu, Daejeon, Republic of Korea

Sangkeun Jung†
hugmanskj@gmail.com
Chungnam National University
Yuseong-gu, Daejeon, Republic of Korea

Riwoo Chung
riwoo.chung@kt.com
KT Corporation
Seocho-gu, Seoul, Republic of Korea

Hwijung Ryu
hwijung.ryu@kt.com
KT Corporation
Seocho-gu, Seoul, Republic of Korea

Du-seong Chang
dschang@kt.com
KT Corporation
Seocho-gu, Seoul, Republic of Korea

ABSTRACT

Summarization is one of the important tasks of natural language processing used to distill information. Recently, the sequence-to-sequence method was applied, in a general manner, to summarization tasks. The problem is that a large amount of information must be pre-trained for a specific domain, and information other than input statements cannot be utilized. To compensate for this shortcoming, controllable summarization has recently been in the spotlight. We introduced three properties into controllable summarization: 1) a new human-machine communication input format, 2) a robust constraint-sensitive summarization method for these formats, and 3) a practical interactive summarization interface available to the user. Experiments on the Wizard-of-Wikipedia dataset show that applying this input format and the constraint-sensitive method enhances summarization performance compared to the typical method. A user study shows that the interactive summarization interface is practical and that participants are evaluating it positively.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies** → **Natural language generation**;

KEYWORDS

Dialogue summarization, text tagging, constraint-sensitive generation, neural networks

ACM Reference Format:
Jeesu Jung, Hyein Seo, Sangkeun Jung, Riwoo Chung, Hwijung Ryu, and Du-seong Chang. 2023. Interactive User Interface for Dialogue Summarization. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3581641.3584057

1 INTRODUCTION

Summarization is one of the most important tasks in natural language processing. It is the process of distilling important information and generating summarized text from documents. In the case of summarization using neural network structures, the summary proceeds sequence-to-sequence. The sequence-to-sequence method has the advantage of being able to learn building a summarization system by only learning data on the domain itself. However, it has the disadvantage of having to learn a large amount of data to acquire domain knowledge. In the case of abstract summarization, this learned model may have a problem of information distortion [Huang et al. 2021]. In addition, except for the input statement, it is hard for the user to interfere with the output statement.

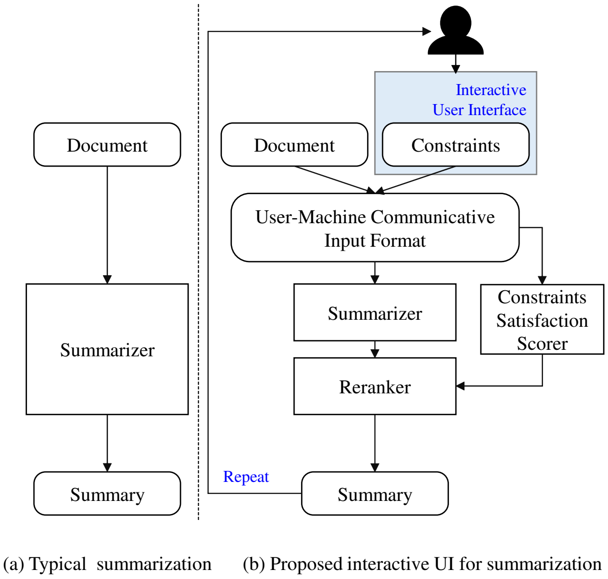
To address these shortcomings, researchers have adopted a variety of approaches, such as question answering [Nan et al. 2021b], knowledge graph [Zhu et al. 2021], and constructive learning [Liu et al. 2021]. As part of the solution to this problem, interest in controllable summarization methods has recently increased beyond traditional summarization. When user controllable elements are added, it is possible to create a new summary sentence with various syntaxes and contexts according to the user's request.

The summarization technique pursued in this study utilizes the following three properties for communication to control the information between the user and artificial intelligence (AI). 1) *User-machine communicative input format* is defined to add controllable information. An intermediate format that is easy to read, both human and machine, writable, and expandable, is required. 2) *Constraint-sensitive summarization* methods are employed. User requirements must somehow be appropriately added in text summarization frameworks. 3) *Interactive summarization interface* is introduced to help with the communication of humans and AI, enabling interactive co-creation with user requirements. This study introduces a novel controllable summarization framework and user interface to address these problems. Figure 1 shows the difference between the existing method and ours.

For the user-machine communicative input format, a **constraint markup language (CML)** was newly proposed and developed for the constraint description of the original dialog. To highlight the important parts, we used this markup language. As demonstrated by a hyper text markup language (HTML), a markup language has many advantages in adding semantic and constraint information to the dialog. The markup language form can minimize the distortion

*Both authors contributed equally to this research.
†Correspondence Author

IUI '23, March 27–31, 2023, Sydney, NSW, Australia
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0106-1/23/03.
https://doi.org/10.1145/3581641.3584057



(a) Typical summarization (b) Proposed interactive UI for summarization

Figure 1: (a) the typical sequence-to-sequence summarization process, (b) our interactive human-guided summarization process.

of information and provide additional information to text in a form understandable to both humans and machines.

In this research, to address constraint-sensitive summarization, we employed *constraint-sensitive training* as well as *generation-and-reranking* approaches. We had to create a new summarizer, sensitive to our new input forms, where critical information was tagged in CML format. To generate the optimized summary, multiple summaries were generated and the summary closest to the information the user wanted was adopted.

To achieve an interactive summarization interface, a user interface was also provided to make the generation method proposed in this paper easier for users. The user can mark the key information to be summarized through the UI and receive a summary in which the information acts as a control element. Through this co-creation process, the user can generate a summary sentence easily, and the AI participates as an assistant.

Massive experiments were conducted in relation to specific food domain data of the Wizard of Wikipedia (WoW) [Dinan et al. 2018] dataset to verify the effectiveness of CML and the proposed summarization frameworks. Our method proved to be better than a summarizer without CML annotated dialogue. in measuring Bilingual Evaluation Understudy(BLEU) [Papineni et al. 2002], Recall-Oriented Understudy for Gisting Evaluation(ROUGE) [Lin 2004], and BERTScore [Zhang* et al. 2020]. Specifically, our proposed method performed better on baselines up to 9.87 measured by BLEU-4gram, 10.24 by ROUGE-L, and 1.83 by BERTScore. In the user study, the participants judged our interactive summarization interface as useful.

In this paper, Section 2 discusses previous studies conducted in this area. Sections 3 and 4 describe the detailed structure of the proposed methodology, and Sections 5 and 6 describe the data used and experimental settings for the neural network. Section 7 describes the user experience and user evaluation of the interface. Section 8 describes the result of the experiments. Finally, Section 9 concludes the entire paper.

2 RELATED WORKS

2.1 Human-AI Collaborative Generation

Recently, the performance of deep generative neural network models has increased significantly. This development has led to practical research on the benefits humans can obtain using these systems. Studies are being conducted in various fields in which humans and systems proceed with generation together using this generation model. There are many studies on co-generation, such as image-oriented generation [Davis et al. 2016; Karimi et al. 2019; Oh et al. 2018], video game content creation [Guzdial et al. 2019], and design generation [Koch et al. 2019]. Voice may be generated for input such as text or converted into text when voice is the input. In addition, various deep learning-based generation systems are being studied for user convenience, such as generating music [Huang et al. 2020; Louie et al. 2020; Suh et al. 2021] for inputs that meet certain conditions or automatically generating meeting minutes [Liu et al. 2020]. The commonality of these systems is that the user takes the initiative to use the AI system for support. This study was conducted with these points in mind.

2.2 Human-AI Collaborative Text Generation

One of the main input and output types utilized by AI systems is text, as text is an objective input form that includes human intention and can be handled easily, not being limited to the surrounding environment. As such, many studies on text-text generation are being conducted. When the user inputs text, the intention is to obtain the desired result. When a specific keyword is an input, it may be combined to generate a sentence, or when a user inputs a word through an input device, multiple words that are likely to be used can be generated and recommended next. Gmail's automatic reply recommendation system [Kannan et al. 2016] and Google Translator [Wu et al. 2016] are common commercial applications.

2.3 Encoder/Decoder-based Neural Summarization

AI is used especially in summary systems for various texts. There are two main types of summary systems: 1) extractive summarization that extracts key sentences and proceeds with a simple concatenation and 2) abstractive summarization that generates a new summary for the entire set of sentences. These summarization systems are being studied in multiple data domains. They are used not only for summaries of common news articles [Gupta et al. 2022], but also for conversation summaries [Zhong et al. 2022] as well as medical document or dialogue summaries [DeYoung et al. 2021; Yadav et al. 2022]. Among them, systems that provide user summary results are also commercialized.

2.4 Language Model-based Applications

As the size of the deep learning model gradually increases, large-scale language models [Brown et al. 2020; Raffel et al. 2020; Shoeybi et al. 2019] are being developed. As data on colloquial text increases due to the commercialization of chat and messenger, various tasks on spoken languages are being studied, such as automatically responding to chat, [Chao and Lane 2019; Oluwatobi and Mueller 2020]. In particular, an interface in the form of generating answers, as if talking to a user, has been studied [Shuster et al. 2022; Smith and Dragone 2022].

2.5 Constraint-based Text Generation

The constraint can be defined differently according to the characteristics of various texts. Both syntax and semantic characteristics can be referred to, as defined and used by many researchers. For example, there is a study using a word that must be included [Fan et al. 2017] or selecting a style such as a colloquial or written language [Ficler and Goldberg 2017] or alternatively using generated text length [Saito et al. 2020; Takeno et al. 2017] as a constraint.

Rendering constraint as input is the simplest form of applying a constraint to a neural network. For training a controllable neural network, one or multiple constraints are appended to the header or the footer as input to the neural network [Gupta et al. 2020; Su et al. 2021a; Takeno et al. 2017].

As an automatic sentence generation system using constraints, [Vig et al. 2021] provides multiple summaries on the document. This system shows the novel words in the summary that do not appear in the source document, but are semantically similar to a token in the source document, and stopwords that are not used in the matching algorithms. Although the system is similar to ours in that it provides information about word distortion common in abstract summarization, our system is different in that it produces new summaries by newly manipulating distorted information.

3 USER GUIDED DIALOGUE SUMMARIZATION WITH INTERACTIVE INTERFACE

The interactive summarization pursued in this study has the following characteristics: 1) the definition of the type of summary constraint meets the needs of the machine-user, 2) the definition of the intermediate format is comprehensible to both users and machines, 3) the improved sequence-to-sequence summary method can react sensitively and accurately to these constraints, and 4) the user interfaces are in the user-available form of the controllable summarization framework.

3.1 Controllable Text Generation

When generating sentences using a transformer-based encoder-decoder model [Vaswani et al. 2017], it is not easy to provide additional important information to the model until the output for a particular input is obtained. There can be various criteria for such information requiring additional inputs.

Typical abstractive summarization provides a more natural context than extractive summarization; however, the problem of the possible increase of information distortion arises [Nan et al. 2021a]. To address this problem, we have defined the concept of *information consistency* that takes the information that users want to keep as is.

In this study, we considered the various constraints the user would generally want to be reflected in summary. The important challenge of summarization is the consistency of information. To maintain information consistency, this study aims to respect the information that users want to maintain.

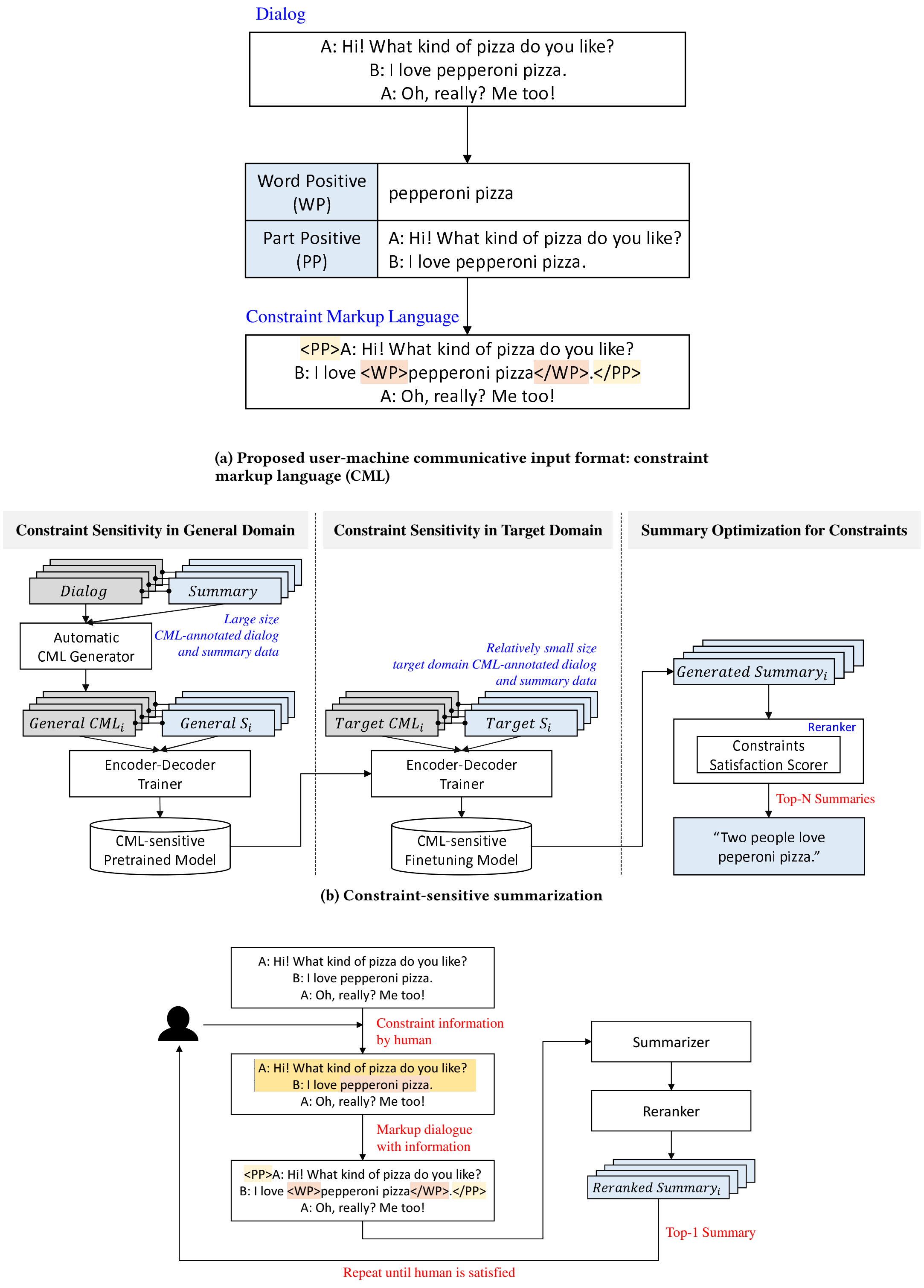
For this goal, the keywords and key phrases described in the summary are defined as constraints.

- **Word Positive (WP)**: Essential keywords to include in the summary without distorting the sentence.
- **Part Positive (PP)**: Key topic sentences or phrases in the summary that is possible to change in a sentence as long as the information is maintained.

3.2 Constraint Markup Language

An input format, jointly understandable by machine and user, is significant and challenging to create. For more user-desirable summarization, we considered an input form that could express the important information in the original text. Several researchers have already devised such a method [Gupta et al. 2021; Su et al. 2021b], but most are arranged with the location in front of the original text. We propose a novel way for putting together the context location information that is *readable* both by humans and AI, *human-writable* and *expandable*.

In this study, we propose CML as a control interface. CML is the constraint description format that uses *markup* tags to express semantic and syntax constraints, which guide summarization modules on how to generate summaries. The markup tags cover a wide range - characters, words, sentences, and entire documents, among others. The tag is expressed on the original document as a format of <tag></tag> just like HTML. This markup language is easy for humans to read/write and can be used as a parser or as input to a machine.


Dialog
A: Hi! What kind of pizza do you like?
B: I love pepperoni pizza.
A: Oh, really? Me too!
Word Positive (WP)
pepperoni pizza
Part Positive (PP)
A: Hi! What kind of pizza do you like?
B: I love pepperoni pizza.
Constraint Markup Language
<PP>A: Hi! What kind of pizza do you like?
B: I love <WP>pepperoni pizza</WP>.</PP>
A: Oh, really? Me too!
Constraint Sensitivity in General Domain
Constraint Sensitivity in Target Domain
Summary Optimization for Constraints
Dialog
Summary
Large size CML-annotated dialog and summary data
Automatic CML Generator
General CML_i
General S_i
Encoder-Decoder Trainer
CML-sensitive Pretrained Model
Relatively small size target domain CML-annotated dialog and summary data
Target CML_i
Target S_i
Encoder-Decoder Trainer
CML-sensitive Finetuning Model
Generated Summary_i
Reranker
Constraints Satisfaction Scorer
Top-N Summaries
"Two people love peperoni pizza."
A: Hi! What kind of pizza do you like?
B: I love pepperoni pizza.
A: Oh, really? Me too!
Constraint information by human
A: Hi! What kind of pizza do you like?
B: I love pepperoni pizza.
A: Oh, really? Me too!
Markup dialogue with information
<PP>A: Hi! What kind of pizza do you like?
B: I love <WP>pepperoni pizza</WP>.</PP>
A: Oh, really? Me too!
Summarizer
Reranker
Reranked Summary_i
Top-1 Summary
Repeat until human is satisfied

(a) Proposed user-machine communicative input format: constraint markup language (CML)

(b) Constraint-sensitive summarization

(c) Interactive summarization interface

Figure 2: The overall architecture and working flow of the proposed summarization framework. An interactive dialogue interface for participants can be viewed here as part of the study.

We further introduce two tags - <WP> and <PP> - for word-level and part-level (multi-words), respectively, to describe key constraints for summarization. The Word Positive (<WP>) tag specifies which word should be contained in the output summary, and Part Positive(<PP>) tag specifies which part of the document should be generated in the summary. Normally, the <PP> tag covers sentence-level information.

3.3 Constraint-Sensitive Summarization Method

For the controllable generation task, a new summarizer, sensitive to the new input forms had to be created with critical information tagged in CML format. This section describes the training, generation, and evaluation methods used to train the constraint in the neural network using pre-training and fine-tuning input format with CML. The sentences are generated from the decoder cover, given the constraints, but the results may not be coherent with the given semantics since the decoder tends to choose the best, based on language model probabilities. To generate CML-consistent results, reranking is performed after generation steps. The overall framework is illustrated in Figure 2.

3.3.1 Constraint-Sensitive on General Domain. The input format proposed in this study is a form with many advantages, but it is a new input format. Therefore, the performance of typical language models is not sufficient for CML. As shown in Figure 2b, the CML-sensitive model was pre-trained, using a transformer-based encoder-decoder architecture as a trainer. The trainer inherits a typical sequence-to-sequence framework, encodes the source CML annotated dialogue as input, and then generates the target summary S with the decoder.

The operation of the summarizer is described in the CML annotated dialogue along with the S, D= $(\mathrm{CML}_1, \mathrm{S}_1)...(\mathrm{CML}_{|\mathrm{D}|}, \mathrm{S}_{|\mathrm{D}|})$ where each instance is a <CML_i, S_i> pair. During the training phase, model parameters are trained to maximize the log-likelihood of the outputs in a parallel training corpus (D):

$$\sum_{(CML,S)\in D} logp(S|CML,\theta) \tag{1}$$

where θ is the model parameter. The decoder generates an abundance of sentences given the CML annotated dialogue.

Automatic CML Generator. We created a system to automatically collect keywords and key phrases based on the answer summary to structure data in CML form. In the case of keywords, Named Entity Recognition (NER) was used to take an intersection between words collected from the summary and the set of words collected from the original dialog. In the case of key phrases, the sentence, with the longest sequence matching, was collected by matching the summary and the original dialog. Keywords and key phrases collected in this process were tagged as <WP> and <PP>, respectively.

3.3.2 Constraint-sensitive on Target Domain. The CML-sensitive pre-trained model was trained on the overall dialogue corpus. After that, fine-tuning was undertaken to learn more specific words or essential phrases that appear for a specific domain. In this study, we fine-tuned the model using CML annotated dialogues and summaries in the target domain. The training method and objective function of the CML fine-tuning model are the same as those of the CML-sensitive pre-trained model.

3.3.3 Summary Optimization for Constraints. To provide an optimized summary including the information desired by the user, the summary was selected through additional algorithms in this study. To achieve controllability, we adopted a *generation-and-reranking* approach such that, 1) a number of summaries are generated, 2) information satisfaction for the summary is evaluated, and 3) the final summary is selected after reranking according to the evaluated score. This method generates a summary that is most consistent with the information desired by the user.

In the generation step, given an input sequence CML annotated dialogue, our model generates multiple summaries that serve as weakly satisfied CML constraints. During the decoding process, the decoder adopts a beam search strategy of the current best hypothesis at each time step.

In this work, we propose new constraint evaluation metrics that satisfy the <WP> and <PP> tags in CML for controllable dialogue summarization tasks: **WP-Accuracy** and **PP-ROUGE-L**.

- **WP-Accuracy** computes the average accuracy of whether the <WP> tags exist in the generated summaries. In evaluating the exact value matching, if <WP> tags in the CML annotated dialog were reflected in generated summary, we were judged to be True, and if not, False.
- **PP-ROUGE-L** calculates the ROUGE-L score between generated summaries and <PP> annotated dialogue. More detailed descriptions of the ROUGE-L score can be seen in Section 6.2.

In the reranking step, the reranker selects the top-N summaries based on the WP-Accuracy or PP-ROUGE-L score. The appropriate summary may be selected from methods that reflect only WP-Accuracy, PP-ROUGE-L, or the sum of the two scores. In this study, we use the sum of the two scores. The multiple generated summaries are then reranked with our new metrics, WP-Accuracy and PP-ROUGE-L. Figure 3 shows a simple example of the process of reranking the generated sentences.

4 UX PATTERNS FOR INTERACTIVE DIALOGUE SUMMARIZATION

In this study, we provide an application tool for the user to easily and quickly transform the input format of the constraint-sensitive neural summarizer. The application tool is configured in the form of a web. Users can then use the following tools. When the user designates the text and important keywords, the service uses a form to output an interactive reflected summary. CML is automatically generated internally in the format of the model input when the user clicks the completion button after displaying the keyword in the text in the form of a drag. Subsequently, a summary result for this is provided to the user. This UX pattern may be repeated until a summary result, satisfactory to the user, is obtained. The total flow is displayed in Figure 4.

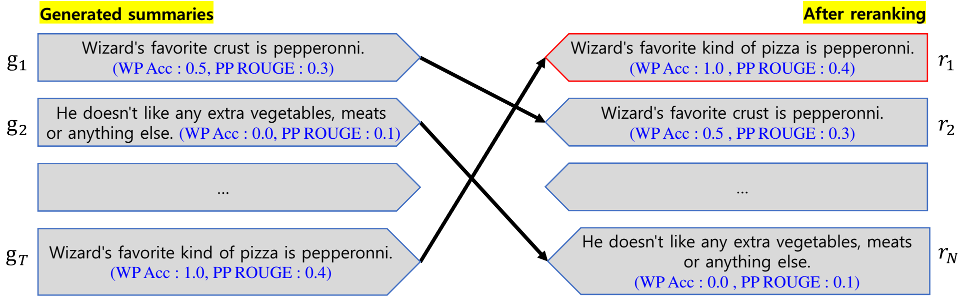

Generated summaries
After reranking
g_1
g_2
g_T
r_1
r_2
r_N
Wizard's favorite crust is pepperonni. (WP Acc : 0.5, PP ROUGE : 0.3)
He doesn't like any extra vegetables, meats or anything else. (WP Acc : 0.0, PP ROUGE : 0.1)
...
Wizard's favorite kind of pizza is pepperonni. (WP Acc : 1.0, PP ROUGE : 0.4)
Wizard's favorite kind of pizza is pepperonni. (WP Acc : 1.0 , PP ROUGE : 0.4)
Wizard's favorite crust is pepperonni. (WP Acc : 0.5 , PP ROUGE : 0.3)
...
He doesn't like any extra vegetables, meats or anything else. (WP Acc : 0.0 , PP ROUGE : 0.1)

Figure 3: Example of reranking generated summaries using constraint evaluation.

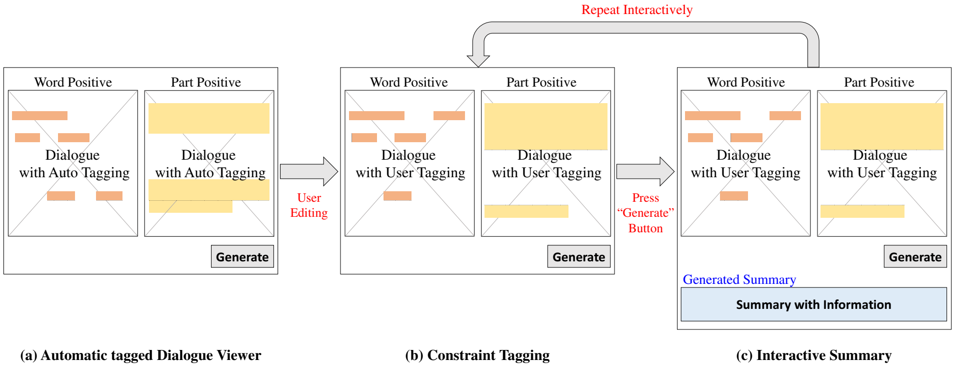


Figure 4: User interface for interactive dialogue summarization. Users can control information by dragging and clicking to generate summaries. This process is repeated until the user is satisfied.

4.1 Automatic tagged Dialogue Viewer

The service provides users with a dialogue with two speakers on a single topic and automatically generated keywords and key phrases. For the demonstration, we first collected the keywords and key phrases on a limited test data set. We collect the summaries automatically generated by 3 different pre-trained models. With those summaries, keywords are extracted from noun intersection using an AI-based part-of-speech tagger. Key phrases are tagged by the longest common subsequence algorithm with the dialogue and answer summary. Users can check the dialogues and highlighted keywords and key phrases, to initially decide on the crucial part.

4.2 Constraint Tagging

Among automatically generated keywords, incorrectly generated keywords may be changed by the user. Keywords irrelevant to the topic can be deleted, or information that must be included can be added. Even if the range is incorrectly tagged, such as when only "Pizza" is tagged among "Pepperoni Pizza", the keyword can be deleted and added again by changing the range.

The tagging interface of keywords and key phrases are provided separately. However, the UX patterns are the same. The user may independently configure keywords and key phrases.

4.3 Interactive Summarization

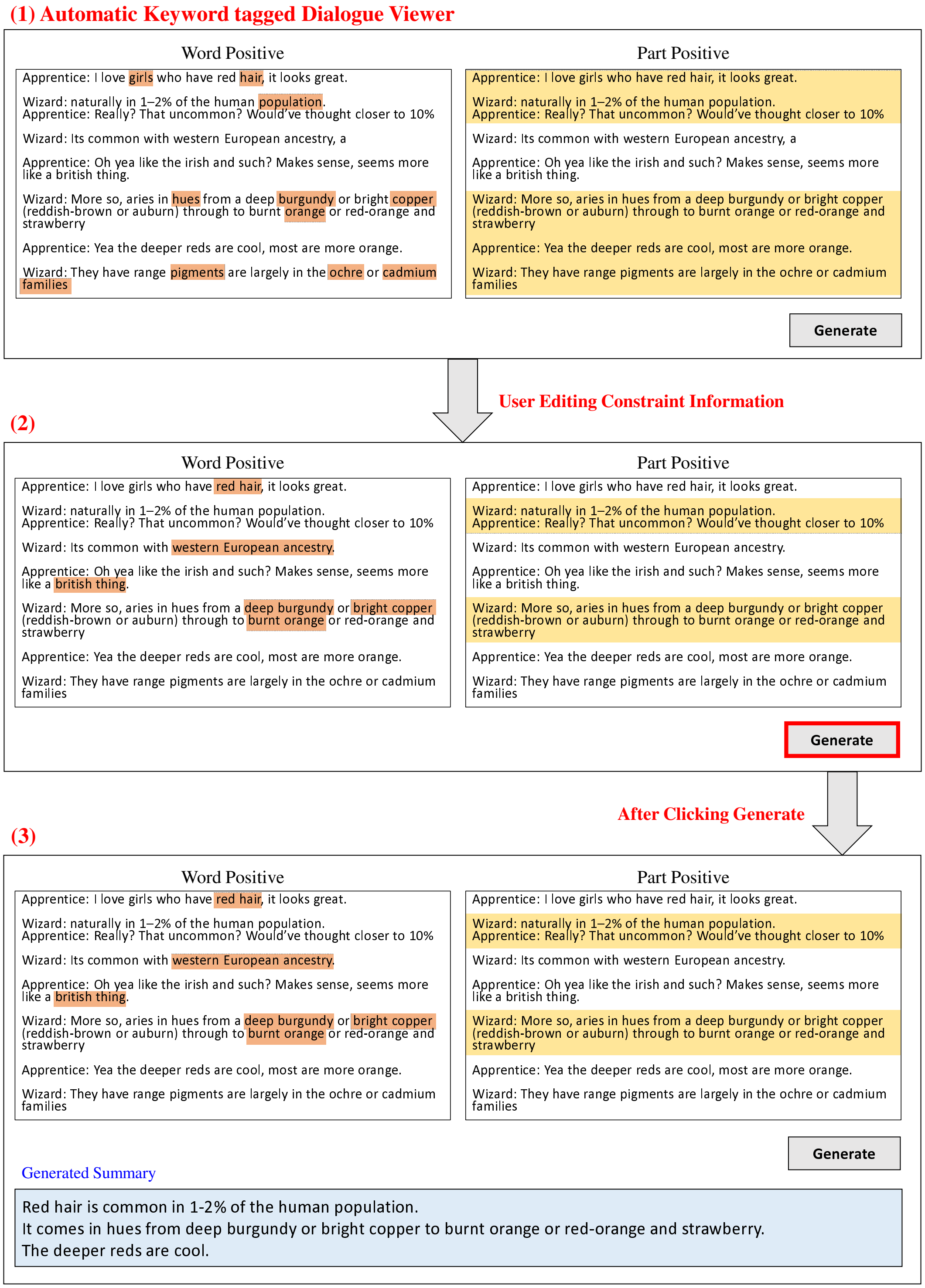
After the user finishes tagging, when the user presses the completion button, the key phrases tagged by the user are reflected in the dialogue in the format of CML. The CML is entered as input into the model, and the model generates summaries reflecting the key phrases. The user can repeat the above process until the generated summary is satisfactory. Figure 5 describes this entire process with the example dialogue.

5 DATASETS

This section introduces the datasets used for CML-sensitive pre-trained models and fine-tuning models. For training, the dialogue-summary pairs were used. To demonstrate the effectiveness of our CML methods on summarization datasets, we chose SAMSum [Gliwa et al. 2019] and WoW. Statistics of dialogue datasets are shown in Table 1.

5.1 Pre-training Datasets

SAMSum is a dialogue summary dataset that contains dialogues from real life. Although the data summaries provide accurate information about the dialogue, these are only short dialogues. WoW is a rich dialogue dataset in which two speakers talk about a specific topic, but do not provide summaries.



Apprentice: I love girls who have red hair, it looks great.

Wizard: naturally in 1–2% of the human population.
Apprentice: Really? That uncommon? Would’ve thought closer to 10%

Wizard: Its common with western European ancestry, a

Apprentice: Oh yea like the irish and such? Makes sense, seems more like a british thing.

Wizard: More so, aries in hues from a deep burgundy or bright copper (reddish-brown or auburn) through to burnt orange or red-orange and strawberry

Apprentice: Yea the deeper reds are cool, most are more orange.

Wizard: They have range pigments are largely in the ochre or cadmium families

Apprentice: I love girls who have red hair, it looks great.

Wizard: naturally in 1–2% of the human population.
Apprentice: Really? That uncommon? Would’ve thought closer to 10%

Wizard: Its common with western European ancestry, a

Apprentice: Oh yea like the irish and such? Makes sense, seems more like a british thing.

Wizard: More so, aries in hues from a deep burgundy or bright copper (reddish-brown or auburn) through to burnt orange or red-orange and strawberry

Apprentice: Yea the deeper reds are cool, most are more orange.

Wizard: They have range pigments are largely in the ochre or cadmium families

Apprentice: I love girls who have red hair, it looks great.

Wizard: naturally in 1–2% of the human population.
Apprentice: Really? That uncommon? Would’ve thought closer to 10%

Wizard: Its common with western European ancestry.

Apprentice: Oh yea like the irish and such? Makes sense, seems more like a british thing.

Wizard: More so, aries in hues from a deep burgundy or bright copper (reddish-brown or auburn) through to burnt orange or red-orange and strawberry

Apprentice: Yea the deeper reds are cool, most are more orange.

Wizard: They have range pigments are largely in the ochre or cadmium families

Apprentice: I love girls who have red hair, it looks great.

Wizard: naturally in 1–2% of the human population.
Apprentice: Really? That uncommon? Would’ve thought closer to 10%

Wizard: Its common with western European ancestry.

Apprentice: Oh yea like the irish and such? Makes sense, seems more like a british thing.

Wizard: More so, aries in hues from a deep burgundy or bright copper (reddish-brown or auburn) through to burnt orange or red-orange and strawberry

Apprentice: Yea the deeper reds are cool, most are more orange.

Wizard: They have range pigments are largely in the ochre or cadmium families

Apprentice: I love girls who have red hair, it looks great.

Wizard: naturally in 1–2% of the human population.
Apprentice: Really? That uncommon? Would’ve thought closer to 10%

Wizard: Its common with western European ancestry.

Apprentice: Oh yea like the irish and such? Makes sense, seems more like a british thing.

Wizard: More so, aries in hues from a deep burgundy or bright copper (reddish-brown or auburn) through to burnt orange or red-orange and strawberry

Apprentice: Yea the deeper reds are cool, most are more orange.

Wizard: They have range pigments are largely in the ochre or cadmium families

Apprentice: I love girls who have red hair, it looks great.

Wizard: naturally in 1–2% of the human population.
Apprentice: Really? That uncommon? Would’ve thought closer to 10%

Wizard: Its common with western European ancestry.

Apprentice: Oh yea like the irish and such? Makes sense, seems more like a british thing.

Wizard: More so, aries in hues from a deep burgundy or bright copper (reddish-brown or auburn) through to burnt orange or red-orange and strawberry

Apprentice: Yea the deeper reds are cool, most are more orange.

Wizard: They have range pigments are largely in the ochre or cadmium families

Red hair is common in 1-2% of the human population.
It comes in hues from deep burgundy or bright copper to burnt orange or red-orange and strawberry.
The deeper reds are cool.

Figure 5: The interactive summarization interface shows how a user can select the controllable part of the Dialogue: (1) viewer displays the important part from the results of automatic detection (2) user can drag or click to edit the information (3) user can then click on the ‘Generate’ button. The demonstration is available at http://168.188.125.27:20063/

			# Dial.	# PP	# Words (PP)	# WP
Pre-training	SAMSum	train	15,550	1.76	187.91	5.85
	WoW	train	19,304	4.19	402.07	15.38
Fine-tuning	WoW-food	train	719	4.27	385.72	15.34
		test	152	4.34	402.16	15.03

Table 1: Statistics of the used dialogue summary datasets, including the total number of dialogues (# Dial.), the average number of Part Positive (PP) sentences, the number of words in PP sentences, and the average number of WP tags per sentence.

Hyperparameter	Pre-training	Fine-tuning
Models	T5-small, BART-base	T5-small, BART-base
Datasets	SAMSum + WoW	WoW-food
CML Tags	WP, PP	WP, PP
Random seed	42	{0,500,1000,1500,2000}
Learning rate	1e-4	1e-4
Batch size	64	32
Epochs	500	200
Optimizer	Adam	Adam
Weight decay	0.01	0.01
Maximum Sequence Length	512	512

Table 2: Experimental setup for hyperparameters.

So we built summaries automatically, using summarization models: text-to-text transfer transformer (T5)[1] [Raffel et al. 2019], bidirectional and auto-regressive transformer (BART)[2] [Lewis et al. 2019], and BART-sum[3] models. We used a summary sentence with the highest proportion of positive words which consists of a noun set that is an intersection of the dialogue and summary. The selected summary ratio of each model was T5: 3.0%, BART: 61.6%, BART-sum: 35.2%.

5.2 Fine-tuning Datasets

As data for fine-tuning, we wanted to build a dialogue dataset for the specific target domain from WoW data. For this purpose, among 22,311 total topics in the WoW dataset, the data were filtered and used for 29 topics related to food. We defined the filtered data as the WoW-food dataset, whereby a training and test set were constructed by dividing in a ratio of 8:2.

6 EXPERIMENTAL SETTINGS

6.1 Implementation Details

6.1.1 Constraint-sensitive Summarizer Details. Our approach can adopt all generation models which can generate sentences corresponding to a given input sequence. The transformer-based encoder-decoder language models were selected for the constraint-sensitive summarization: T5 and BART. T5 is an encoder-decoder architecture using the text-to-text format as input to generate target text. BART is an encoder-decoder transformer model pre-trained on a large corpus using a denoising auto-encoder task that first uses a noise-added source text as input and subsequently uses a language model for reconstructing the original text by predicting the true replacement of corrupted tokens. The T5 and BART models show excellent performance when used for natural language generation(NLG) tasks. For pre-training and fine-tuning the CML-sensitive summarizer, the CML annotated dialogue was fed to the encoder, and the decoder produced summaries satisfying the constraint of the input. Figure 6 shows the overall input and output design of the summarizer. For implementation, we used T5-small[4] and BART-base[5].

6.1.2 Decoding. When decoding at test time, beam search was used to generate candidate summaries with a size of {1,3,5,10,30,50,100}. During the test process, the maximum sequence length was set to 200.

6.1.3 Implementation Details. All our models were trained on NVIDIA A100, using the Huggingface Pytorch transformers[6] package [Wolf et al. 2019]. The experiments' hyperparameters are shown in Table 2. In the pre-training stage, the combined datasets of SAMSum and WoW were used and in the fine-tuning stage, the WoW-food dataset. During the fine-tuning, the model was trained using WoW-food with five different seeds selected from {0,500,1000,1500,2000}.

6.2 Evaluation Metrics

All evaluation metrics range from zero to one with closer to one representing more accuracy and similarity between the reference and generated summaries.

[1] https://huggingface.co/bhuvaneswari/t5-small-text_summarization

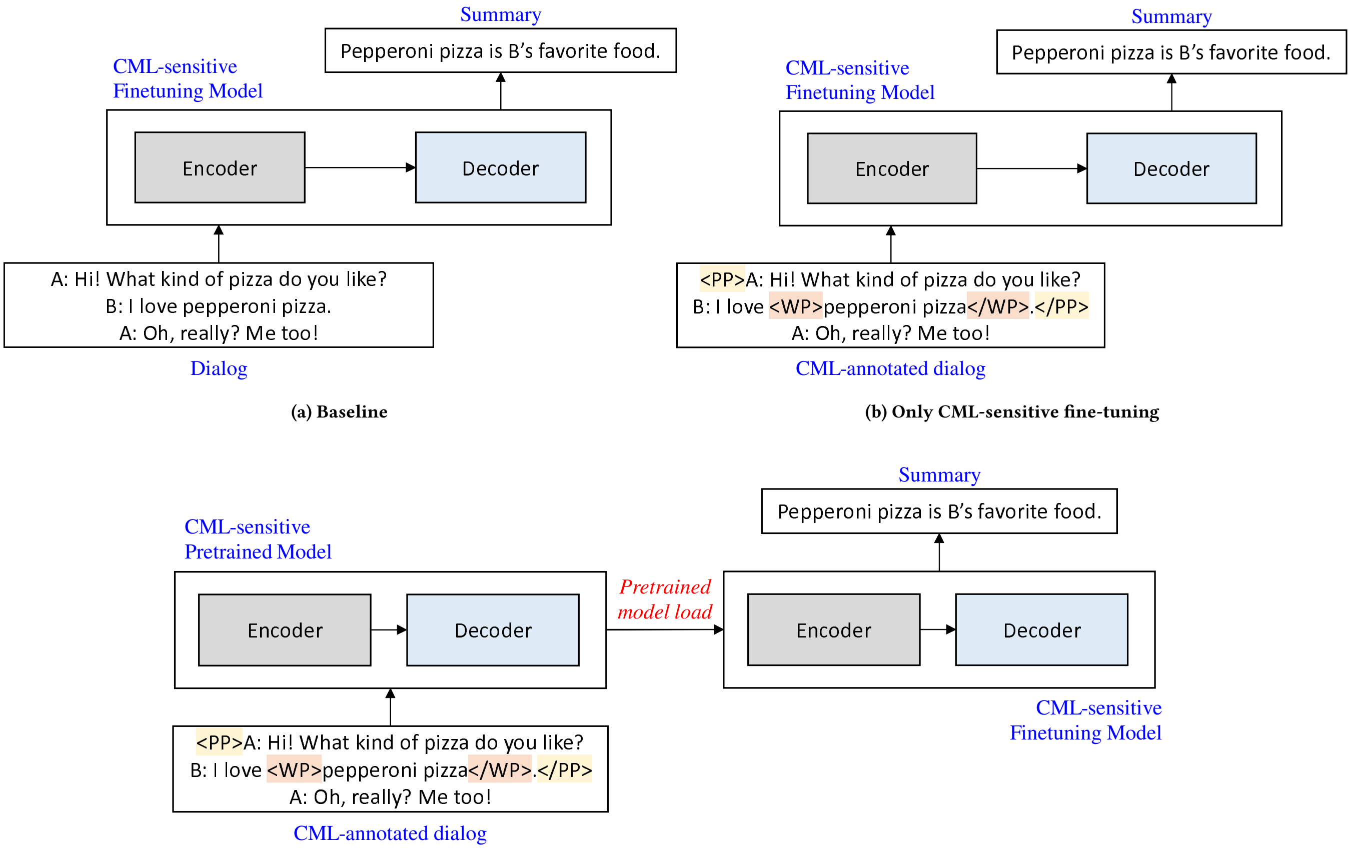
[2] https://huggingface.co/slauw87/bart_summarisation

[3] https://huggingface.co/philschmid/bart-large-cnn-samsum

[4] https://huggingface.co/t5-small

[5] https://huggingface.co/facebook/bart-base

[6] https://github.com/huggingface/transformers


Summary
Pepperoni pizza is B's favorite food.
CML-sensitive
Finetuning Model
Encoder
Decoder
A: Hi! What kind of pizza do you like?
B: I love pepperoni pizza.
A: Oh, really? Me too!
Dialog
Summary
Pepperoni pizza is B's favorite food.
CML-sensitive
Finetuning Model
Encoder
Decoder
<PP>A: Hi! What kind of pizza do you like?
B: I love <WP>pepperoni pizza</WP>.</PP>
A: Oh, really? Me too!
CML-annotated dialog
CML-sensitive
Pretrained Model
Encoder
Decoder
<PP>A: Hi! What kind of pizza do you like?
B: I love <WP>pepperoni pizza</WP>.</PP>
A: Oh, really? Me too!
CML-annotated dialog
Pretrained
model load
Summary
Pepperoni pizza is B's favorite food.
Encoder
Decoder
CML-sensitive
Finetuning Model

(a) Baseline

(b) Only CML-sensitive fine-tuning

(c) CML-sensitive pre-training + fine-tuning

Figure 6: The input and output design of CML-sensitive summarizer.

6.2.1 n-gram Overlap-based Metrics. We measured the quality of summarization commonly adopted for evaluating natural language generation system, including BLEU and ROUGE-L, to assess the coherence of generated summaries. BLEU is a corpus-level precision-focused metric that calculates n-gram overlap between the reference and generated summaries and includes a brevity penalty (BP). N is the maximum length of n-grams considered, BLEU-1/2/3/4 measures the overlap of unigrams/bigrams/trigrams/four-grams about single tokens. The BLEU score as follows:

$$BP = e^{min(1 - \frac{len(reference-summary)}{len(generated-summary)}, o)} \tag{2}$$

$$BLEU = BP * exp(\sum_{k=1}^{n} w_k log(p_k)) \tag{3}$$

k is the number of n-grams and w_k is the positive weight summing to one. $len(reference\text{-}summary)$ and $len(generated\text{-}summary)$ refer to the length of the reference and generated summary, respectively. p_k means the n-gram precisions.

Initially, ROUGE is proposed to test the effectiveness of automatic summarization of long texts containing multiple sentences by comparing the overlap of n-grams, word sequences, and word pairs. In this study, we used the ROUGE-L version, which calculates the recall for the longest common subsequence (LCS) between the reference and generated summaries. The LCS is a set of words that occur in two sentences in the same order, but unlike n-grams, these words do not have to be contiguous, i.e., there can be ordered words between the terms of the LCS. Given a reference summary consisting of m words and a generated summary of length n, the ROUGE-L score is calculated as the weighted harmonic mean of R_{LCS} and P_{LCS} as follows:

$$R_{LCS} = \frac{LCS(reference\text{-}summary, generated\text{-}summary)}{m} \tag{4}$$

$$P_{LCS} = \frac{LCS(reference\text{-}summary, generated\text{-}summary)}{n} \tag{5}$$

$$\beta = \frac{P_{LCS}}{R_{LCS}} \tag{6}$$

$$ROUGE\text{-}L = \frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{7}$$

6.2.2 Embedding-based Metric. The n-gram overlap-based evaluation relies on string matching; we cannot identify paraphrases of the word like "expensive" and "costly." Additionally, we used the semantic matching metrics BERTScore. BERTScore is an embedding-based metric that is more robust to changes in the surface forms of the words than BLEU and ROUGE-L. BERTScore measures soft overlap between contextual bidirectional encoder representations from transformer (BERT) [Devlin et al. 2019] embeddings of tokens between the reference and the generated summaries.

6.2.3 Constraint-based Metrics. In addition, we adopted constraint satisfaction evaluation metrics especially designed for the controllable dialogue summarization task to evaluate the quality of <WP> and <PP> tags in CML, WP-Accuracy and PP-ROUGE-L mentioned in Section 3.3.3.

6.3 Model Baseline

We compared our model with the baseline. For the baseline, the original dialogues without CML annotated tags were fed to the model, and the decoder generated a summary. The architecture and training method of baseline (Figure 6a and CML-sensitive fine-tuning models (Figure 6b) are the same; only the input design is different. The baseline was trained with the same experimental settings as those of fine-tuning in Table 2. Web-based CML-based summarizer services that users can access were built, using the CML-sensitive fine-tuning model.

7 USER STUDY

Based on the augmented model introduced earlier, we devised a web-based service that users can access. The following UI service utilization experiment was conducted to verify how helpful these interactive conversation summary services are to actual users.

7.1 Methodology

We recruited 31 participants interested in AI-related systems through a notice posted at our institution. Participants were tested without receiving any other compensation. Information was collected on their gender and age (Figure 7a), English level (Figure 7b), the highest level of education (Figure 7c), knowledge level on AI (Figure 7d), and major (Figure 7e). They were unaware of the content of this paper and dialogue summarization interfaces. We provided the user interface and asked them to use it for 10 minutes to proceed with the summary of the sample dialogues. The experiment was conducted on the following three cases. The user was presented with five different dialogues and generated a summary of them using the appropriate tool to suit the required topic.

- Summaries generated immediately from dialogue without additional input. (Baseline)
- Summaries generated from the CML automatically tagged constraints. (Auto CML)
- Summaries generated from the interactive CML tagged constraints by the user. (Interactive CML)

7.2 Evaluation

The researchers tried to determine the users' reactions to the interactive summarization interface through a survey. An actual interview on the service was conducted with 31 users using a Likert scale of 1-5 points based on the system usability scale(SUS) [Brooke et al. 1996] questionnaire. The questions asked in the usability questionnaire were as follows:

(1) I think that I would like to use this system frequently.

(2) I found that the system is unnecessarily complex and unintuitive.

(3) I thought the system was easy to use.

(4) I think that I would need the support of a technical person to be able to use this method.

(5) I found the various functions in this system well-integrated.

(6) I thought there was too much inconsistency in this system.

(7) I imagine that most people would learn to use this system very quickly.

(8) I found the system very cumbersome to use.

(9) I felt very confident using the system.

(10) I needed to learn a lot of things before I could get going with this system.

8 RESULTS

The purpose of this study was to combine three elements to make it easier for the user to summarize the conversation. The three elements consisted of: 1) an efficient new controllable input form, CML, that both users and machines could understand; 2) a constraint-sensitive summarizer that could respond to this input form; 3) an interactive summarization interface that would allow users to easily co-create summaries with AI.

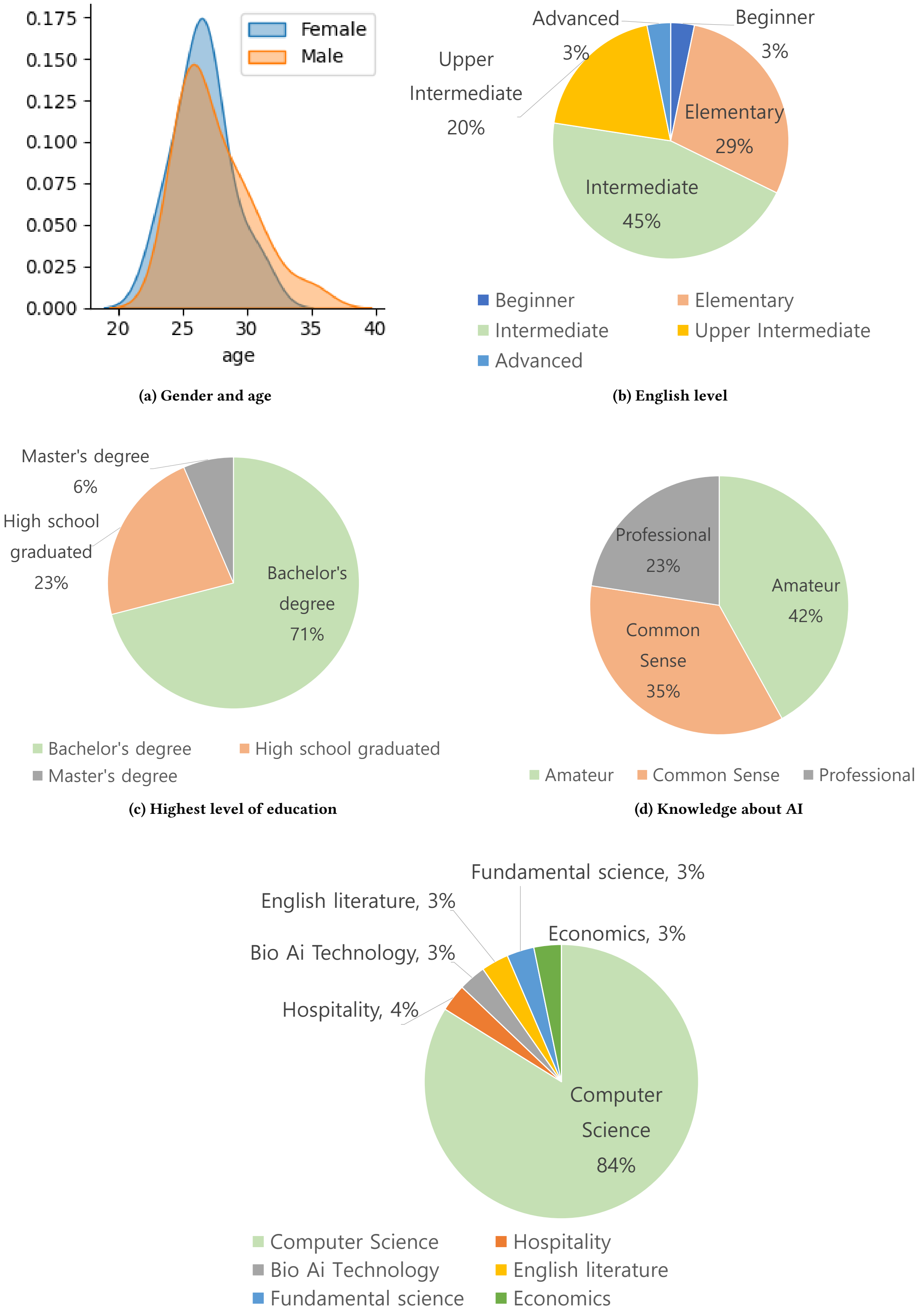
8.1 Effects of Constraint Sensitive Summarization using CML

To verify the effectiveness of the CML annotated dialogues of the proposed method, the model was compared with the baseline. The generated top-1 summary was evaluated after reranking. The performances of BLEU, ROUGE-L, BERTScore, WP-Accuracy, and PP-ROUGE-L are reported as an evaluation of the quality of the generated summaries.

Our experiment involved training on 1) the baseline without CML annotated data, 2) only CML-sensitive fine-tuning, and 3) CML-sensitive pre-training and fine-tuning. Our reranking strategy can be categorized into three groups: (1) SUM: the sum of WP-Accuracy and PP-ROUGE-L scores; (2) WP: only WP-Accuracy; (3) PP: only PP-ROUGE-L.

Table 3 shows the experimental results on the WoW-food test set. Overall, the CML-sensitive models performed better than the the baseline. When using CML-sensitive pre-trained models, all evaluation scores improved significantly, suggesting that the proposed method helps summarizers capture the constraints. The CML-sensitive pre-training + fine-tuning models achieved the best scores. The experimental results for more settings are shown in Appendix A.

Figure 8 and Figure 9 show the average performance of summarization of T5 and BART. We evaluated the summary quality using BLEU-4, ROUGE-L, WP-Accuracy, and PP-ROUGE-L.



(a) Gender and age

(b) English level

(c) Highest level of education

(d) Knowledge about AI

(e) Major

Figure 7: User study participants' information.

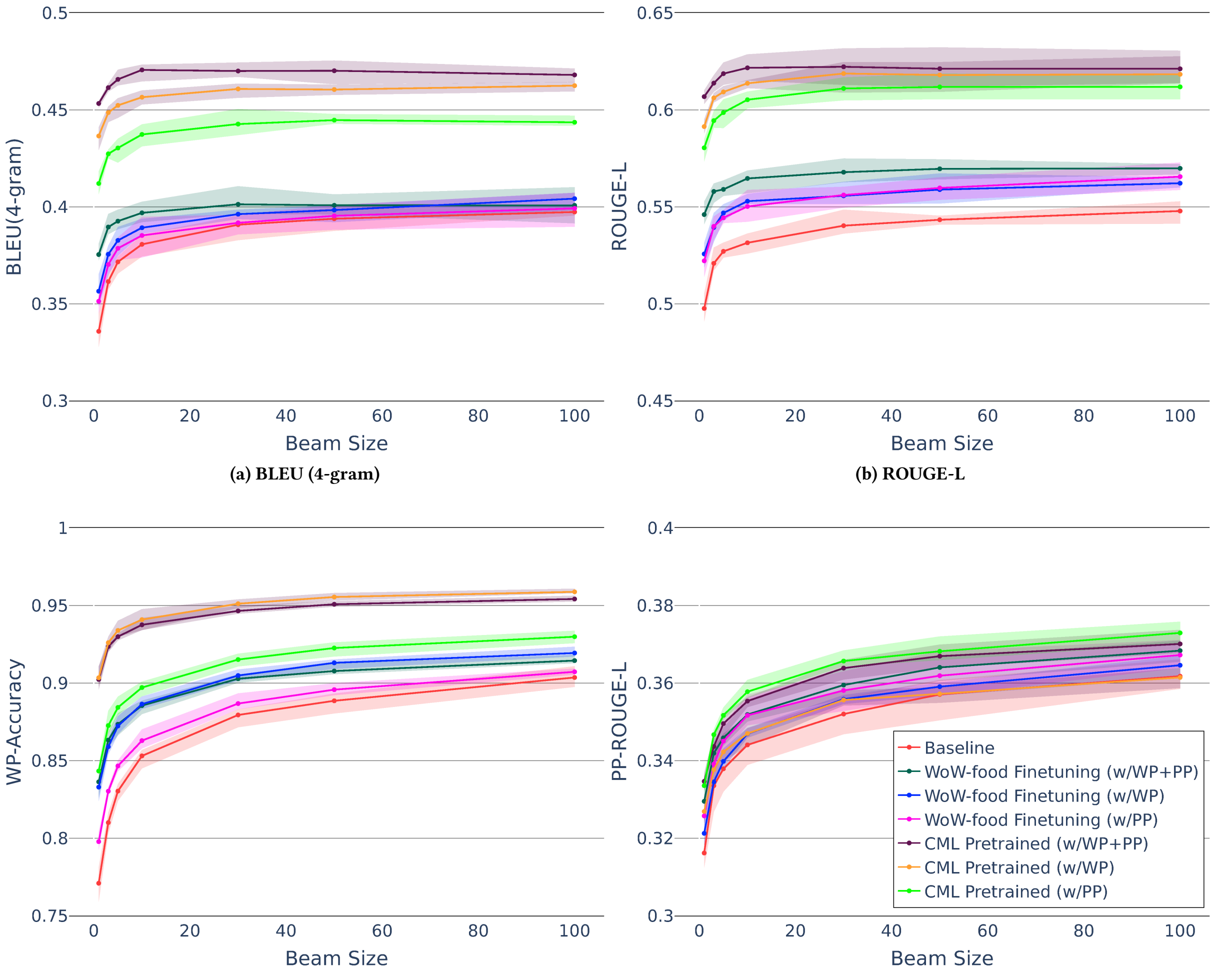
Model	Beam Size	B-1	B-2	B-3	B-4	R-L	BERTScore	WP-Accuracy	PP-ROUGE-L
T5									
Baseline (w/o CML)	1	55.32(±0.4)	45.49(±0.6)	38.79(±0.8)	33.58(±0.9)	49.76(±0.7)	92.17(±0.1)	77.09(±1.1)	31.62(±0.4)
	3	57.64(±0.4)	48.06(±0.4)	41.39(±0.4)	36.15(±0.4)	52.08(±0.6)	92.61(±0.1)	81.00(±0.7)	33.35(±0.5)
	5	58.57(±0.4)	49.12(±0.5)	42.45(±0.5)	37.16(±0.4)	52.70(±0.4)	92.79(±0.1)	83.03(±0.5)	33.79(±0.5)
	10	59.45(±0.6)	50.06(±0.6)	43.39(±0.6)	38.06(±0.6)	53.14(±0.5)	92.95(±0.0)	85.30(±0.6)	34.40(±0.5)
	30	60.33(±0.4)	51.06(±0.6)	44.40(±0.7)	39.08(±0.6)	54.03(±0.6)	93.16(±0.0)	87.94(±0.6)	35.20(±0.5)
	50	60.95(±0.2)	51.60(±0.4)	44.82(±0.5)	39.39(±0.6)	54.33(±0.2)	93.25(±0.0)	88.85(±0.6)	35.71(±0.5)
	100	61.55(±0.2)	52.16(±0.1)	45.29(±0.1)	39.73(±0.2)	54.78(±0.6)	93.31(±0.1)	90.36(±0.7)	36.18(±0.4)
Only CML-sensitive fine-tuning	1	60.30(±1.0)	50.52(±1.0)	43.39(±0.9)	37.77(±0.9)	55.00(±0.7)	93.17(±0.2)	85.92(±1.0)	33.15(±0.4)
	3	62.11(±0.4)	52.58(±0.5)	45.53(±0.6)	39.85(±0.7)	56.67(±0.2)	93.50(±0.1)	88.62(±0.4)	34.73(±0.3)
	5	62.81(±0.5)	53.23(±0.6)	46.15(±0.6)	40.45(±0.6)	56.98(±0.5)	93.62(±0.1)	89.74(±0.5)	35.25(±0.2)
	10	63.24(±0.4)	53.76(±0.5)	46.66(±0.5)	40.91(±0.5)	57.30(±0.3)	93.72(±0.1)	90.89(±0.3)	35.87(±0.4)
	30	63.75(±0.6)	54.25(±0.6)	47.10(±0.4)	41.29(±0.4)	57.51(±0.1)	93.80(±0.1)	92.32(±0.3)	36.74(±0.3)
	50	63.68(±0.4)	54.13(±0.4)	47.00(±0.4)	41.21(±0.3)	57.50(±0.7)	93.83(±0.0)	93.18(±0.5)	37.02(±0.4)
	100	63.88(±0.3)	54.42(±0.5)	47.33(±0.5)	41.59(±0.4)	57.56(±0.6)	93.84(±0.1)	93.93(±0.4)	**37.52(±0.5)**
CML-sensitive pre-training + fine-tuning	1	68.16(±0.2)	59.23(±0.2)	52.30(±0.2)	46.53(±0.2)	61.36(±0.6)	94.44(±0.1)	91.93(±0.1)	33.60(±0.3)
	3	69.04(±0.4)	60.18(±0.5)	53.22(±0.5)	47.43(±0.5)	62.45(±0.2)	94.62(±0.0)	93.63(±0.3)	34.64(±0.1)
	5	69.38(±0.3)	60.50(±0.2)	53.52(±0.1)	47.71(±0.1)	62.83(±0.2)	94.70(±0.0)	94.46(±0.3)	35.06(±0.1)
	10	69.70(±0.2)	**60.78(±0.3)**	**53.76(±0.2)**	**47.93(±0.2)**	**63.38(±0.4)**	**94.78(±0.1)**	95.26(±0.4)	35.48(±0.1)
	30	**69.71(±0.4)**	60.65(±0.5)	53.58(±0.6)	47.72(±0.6)	63.18(±0.8)	94.76(±0.1)	96.19(±0.2)	36.13(±0.2)
	50	69.49(±0.3)	60.41(±0.3)	53.35(±0.4)	47.52(±0.4)	63.08(±0.7)	94.77(±0.1)	96.53(±0.1)	36.50(±0.2)
	100	69.55(±0.4)	60.54(±0.5)	53.52(±0.6)	47.71(±0.6)	63.19(±0.6)	94.77(±0.1)	**96.87(±0.1)**	36.95(±0.2)
BART									
Baseline (w/o CML)	1	55.19(±1.3)	44.90(±1.4)	38.04(±1.3)	32.74(±1.2)	49.19(±0.9)	92.05(±0.2)	74.00(±1.3)	31.38(±0.5)
	3	57.07(±1.5)	47.07(±1.7)	40.27(±1.7)	34.90(±1.5)	51.03(±1.0)	92.37(±0.3)	77.26(±1.1)	32.90(±0.5)
	5	57.85(±1.0)	47.91(±1.2)	41.10(±1.3)	35.71(±1.2)	51.63(±0.7)	92.55(±0.3)	78.59(±0.9)	33.58(±0.4)
	10	58.83(±1.3)	48.91(±1.4)	42.02(±1.4)	36.60(±1.3)	52.23(±1.0)	92.75(±0.2)	80.98(±0.8)	34.19(±0.3)
	30	60.28(±1.2)	50.45(±1.3)	43.49(±1.5)	37.96(±1.4)	53.70(±1.4)	93.02(±0.3)	83.89(±1.0)	35.42(±0.6)
	50	60.86(±0.9)	51.08(±1.3)	44.07(±1.4)	38.48(±1.4)	54.03(±1.0)	93.13(±0.2)	85.06(±1.2)	35.88(±0.4)
	100	61.51(±0.8)	51.76(±1.1)	44.73(±1.2)	39.10(±1.1)	54.51(±1.0)	93.26(±0.2)	86.60(±0.7)	36.39(±0.6)
Only CML-sensitive fine-tuning	1	58.77(±1.5)	48.72(±1.4)	41.62(±1.2)	35.97(±1.1)	53.29(±0.7)	92.76(±0.2)	80.84(±2.2)	31.95(±0.9)
	3	60.74(±1.0)	50.81(±0.8)	43.58(±0.8)	37.83(±0.7)	54.76(±0.5)	93.09(±0.2)	83.44(±1.6)	33.14(±0.3)
	5	61.33(±0.9)	51.41(±0.7)	44.21(±0.5)	38.43(±0.5)	55.34(±0.7)	93.20(±0.2)	84.61(±1.7)	33.73(±0.4)
	10	61.92(±0.9)	51.99(±0.9)	44.79(±0.8)	38.99(±0.8)	55.55(±0.6)	93.28(±0.2)	86.58(±1.0)	34.32(±0.2)
	30	62.75(±0.6)	52.94(±0.7)	45.75(±0.7)	39.92(±0.8)	56.39(±0.7)	93.42(±0.1)	88.89(±0.7)	35.17(±0.2)
	50	63.21(±0.6)	53.33(±0.6)	46.07(±0.6)	40.20(±0.7)	56.76(±0.6)	93.51(±0.1)	89.79(±0.9)	35.61(±0.2)
	100	63.60(±0.7)	53.74(±0.5)	46.53(±0.5)	40.71(±0.5)	57.29(±0.6)	93.58(±0.1)	90.84(±0.9)	35.95(±0.3)
CML-sensitive pre-training + fine-tuning	1	67.67(±0.9)	58.61(±1.0)	51.58(±1.1)	45.74(±1.2)	62.03(±0.7)	94.36(±0.1)	90.06(±0.5)	32.93(±0.6)
	3	68.75(±0.5)	59.87(±0.7)	52.98(±0.8)	47.29(±0.9)	63.40(±0.6)	94.57(±0.1)	91.67(±0.5)	34.13(±0.4)
	5	69.33(±0.7)	60.43(±0.8)	53.54(±0.8)	47.85(±0.9)	63.70(±0.6)	94.68(±0.1)	92.70(±0.3)	34.44(±0.2)
	10	69.66(±0.8)	60.70(±1.0)	53.75(±1.1)	48.02(±1.2)	63.88(±0.7)	94.73(±0.1)	93.53(±0.4)	35.04(±0.4)
	30	70.00(±0.5)	61.06(±0.7)	54.11(±0.7)	48.38(±0.8)	64.13(±0.7)	94.81(±0.1)	94.84(±0.2)	35.95(±0.5)
	50	69.94(±0.7)	60.95(±0.7)	53.98(±0.6)	48.23(±0.6)	64.04(±0.4)	94.78(±0.1)	95.31(±0.3)	36.28(±0.3)
	100	**70.14(±0.6)**	**61.17(±0.6)**	**54.21(±0.6)**	**48.46(±0.8)**	**64.26(±0.7)**	**94.81(±0.1)**	**95.69(±0.2)**	**36.79(±0.3)**

Table 3: Experimental results of automatic metrics on the WoW-food test set. The reranking strategy is SUM (sum of WP-Accuracy and PP-ROUGE-L). For short, B and R refer to BLEU and ROUGE, respectively. The results were averaged over five random runs. The highest numbers are in bold.

8.2 Effects of Interactive User Interface

Figure 10a shows the result of the scoring. Looking at the score distribution of users' survey results, our UI generally scored high on questions 1, 3, 5, 7, and 9, which are positive questions, and low on questions 2, 4, 6, 8, and 10, which are negative questions. This tendency shows that the users generally had positive feelings for our UI. Figure 10b is the score of participant satisfaction with the summary function of the system on a Likert scale of 1-5 points. The median satisfaction score was 4. This is a high score indicating that users judged our system positively.

In addition, we asked participants to choose the most preferred of the three summaries: Baseline, Auto CML, and Interactive CML. This distribution can be seen in Figure 10c with 68% of all users, more than a majority, preferring CML. In particular, interactive CML was the most preferred summarization method. Participants



(a) BLEU (4-gram)

(b) ROUGE-L

(c) WP-Accuracy

(d) PP-ROUGE-L

Figure 8: Average of summarization performances using training model T5 with five different random seeds. Beam size(x-axis) is the number of results used for reranking; w/WP+PP, w/WP, and w/PP refer to the training of the CML corpus with each tag version; w/WP+PP means the model was trained on constraint markup language (CML) using annotated data containing WP and PP tags; w/WP means the CML annotated data used only WP tags; and w/PP means the CML annotated data used only PP tags.

preferred baseline citing "because it's annoying to modify it directly" and "convenient because it's automatically processed." For those who preferred CML cited, "A summary depends on what the word difference alone focuses on, adding and deleting the desired words or sentences, excluding unimportant parts of the entire conversation flow or focusing on what they want" and "When I checked the results of the summary, it came out well similar to the summary I thought." as the reasons.

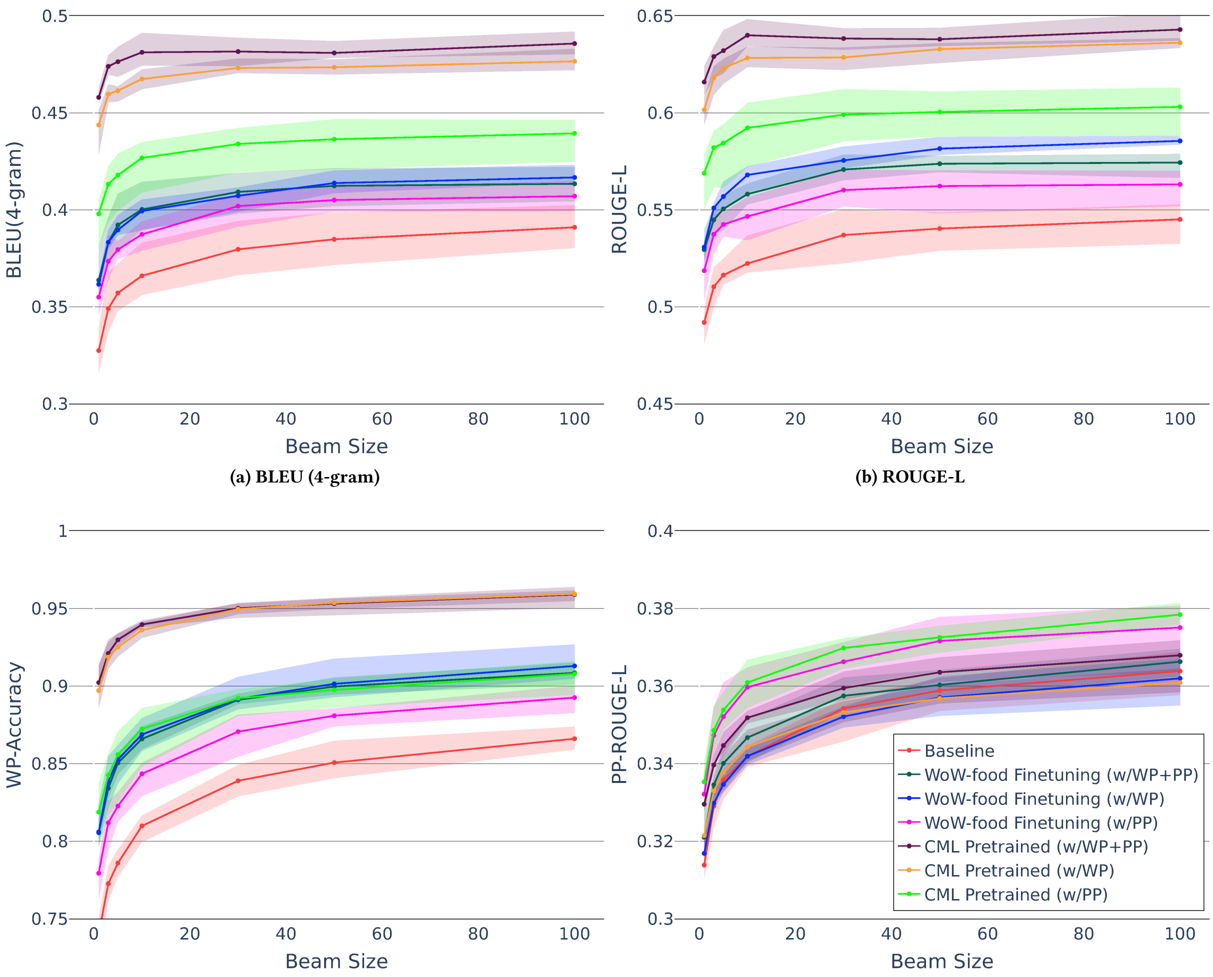
Figure 11a shows the number of submissions made by users made for one original text in the system. 61% of users received satisfactory results with fewer than three submissions. Figure 11b and 11c show the number of modifications made in the system to WP and PP for one original text. In general, for those who were satisfied with Auto CML, few modifications were made, and there were slightly more people making corrections to fewer than three keywords.

8.3 Case Study

A case study was performed to understand the performance of our method better. Table 4 shows the generated summaries for baseline, only CML-sensitive fine-tuning, and CML-sensitive pre-training + fine-tuning in the WoW-food dataset on the T5 model. Comparing the generated summaries, all models could generate fluent and natural sentences. However, the models that adopted the CML-annotated dialogue as input format - only CML-sensitive fine-tuning and CML-sensitive pre-training + fine-tuning, could satisfy more constraints. This confirms our method's efficiency in generating a more controllable and constraint-sensitive summary.

Dialog	**Apprentice**: my favorite food is pizza **Wizard**: pizza is excellent, so glad the Italians passed it down to us, typically a yeasted flatbread with tomato sauce cheese and toppings **Apprentice**: when was the pizza first made? **Wizard**: well the name pizza was 1st recorded in the 10th century in southern Italy **Apprentice**: did Italy invent the pizza or did they get it from someone else **Wizard**: well modern pizza was invented in Naples so I don't think so **Apprentice**: ok, is a deep pan pizza really a pizza :) **Wizard**: yes any kind of crust with toppings is considered pizza, there are even some made without cheese, and they are still considered pizza **Apprentice**: yeah, some are cheese-based and some are tomato sauce-based only **Wizard**: some of the best pizza I have had was on the streets of NY when I was a kid, from the street vendors
CML annotated dialog	<PP><WP>**Apprentice**</WP>: my favorite <WP>food</WP>is <WP>pizza</WP> <WP>**Wizard**</WP>: pizza is excellent, so glad the Italians passed it down to us, typically a yeasted flatbread with tomato sauce cheese and <WP>toppings</WP> **Apprentice**: when was pizza first made?</PP> <PP>**Wizard**: well the name pizza was 1st recorded in the 10th <WP>century</WP> in <WP>southern</WP> <WP>italy</WP> **Apprentice**: did Italy invent the pizza or did they get it from someone else **Wizard**: well modern pizza was invented in <WP>naples</WP> so I don't think so **Apprentice**: ok, is a deep pan pizza really a pizza :) **Wizard**: yes a<WP>NY</WP> kind of <WP>crust</WP> with toppings is considered pizza, there are even some made without cheese and they are still considered pizza **Apprentice**: yeah, some are cheese based and some are tomato sauce-based only **Wizard**: some of the best pizza I have had was on the <WP>streets</WP> of NY when I was a <WP>kid</WP> from the <WP>street</WP> <WP>vendors</WP></PP>
Reference Summary	Wizard's and Apprentice's favorite food is pizza. Pizza was first recorded in the 10th century in Southern Italy. The modern pizza was invented in Naples. The best pizza Wizard ever had was on the streets of NY when he was a kid from the street vendors. Any crust with toppings is considered a pizza.
Model	**Generated Summaries**
Baseline	Apprentice's favorite food is pizza. The name pizza was 1st recorded in the 10th century in Southern Italy. Modern pizza was invented in Naples. Some pizzas are cheese based and others are tomato sauce-based. Apprentice's favorite pizza was on the streets of NY when he was a kid from the street vendors.
Only CML-sensitive fine-tuning	Apprentice's favorite food is pizza. The name pizza was 1st recorded in the 10th century in Southern Italy. Modern pizza was invented in Naples. Wizard had pizza on the streets of NY when he was a kid, from the street vendors. Some of the best pizzas he has had were on the streets of NY when he was a pizza from the street vendors. He doesn't think so.
CML-sensitive pre-training + fine-tuning	Apprentice's favorite food is pizza. The name pizza was first recorded in the 10th century in southern italy. Pizza was invented in Naples. Some of the best pizza Wizard has had was on the streets of NY when he was a kid from the street vendors. A deep pan pizza with toppings is considered pizza, some are cheese based and some are tomato sauce-based.

Table 4: Example results of generated summaries in the WoW-food dataset on the T5 model. The keywords that satisfied the constraints are highlighted in colors. (Best viewed in color)



(a) BLEU (4-gram)

(b) ROUGE-L

(c) WP-Accuracy

(d) PP-ROUGE-L

Figure 9: Average of summarization performances of the trained model with five different random seeds on BART.

9 CONCLUSIONS

In this paper, we introduced the interactive dialogue summarization method. To generate an interactive and information-consistent summary, three properties were defined. We proposed a new user-machine communicative input format called CML, constraint-sensitive summarization methods, and an interactive summarization interface for human-AI co-generation. Experimental results show that the CML and constraint-sensitive models provide useful information, increasing abstractive dialogue summarization performance. In the user study, the interactive summation interface proved effective. In addition, participants found the interactive summarization interface to be more helpful than the baseline which generated in the traditional sequence-to-sequence method.

Our results provide lessons and insights into human and AI co-generation systems and suggest several directions for future research. In particular, we find that the automatic generation of the speaker can be distorted in the dialogue summarization task. We can solve these problems by expanding our methodology. In addition, research will be conducted to expand the techniques introduced in this study on summarizing, to other domains not limited to conversation summary.

ACKNOWLEDGMENTS

This work was supported by Chungnam National University.

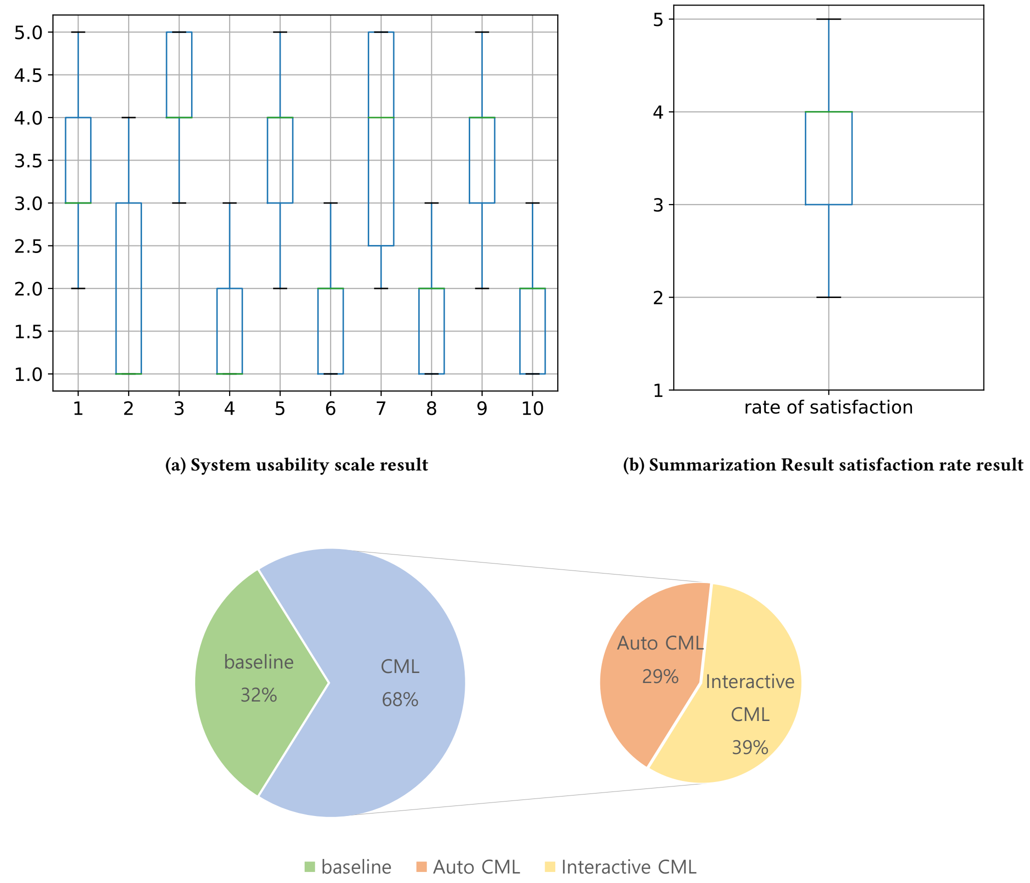
REFERENCES

John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

Guan-Lin Chao and Ian Lane. 2019. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. *arXiv preprint arXiv:1907.03040* (2019).

Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, and Brian Magerko. 2016. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 196–207.



(a) System usability scale result

(b) Summarization Result satisfaction rate result

(c) Preference for summary form

Figure 10: (a) The score distribution of 10 questions of SUS. (b) Participants' preference for summary form (Baseline, Auto CML, Interactive CML).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486* (2021).

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241* (2018).

Angela Fan, David Grangier, and Michael Auli. 2017. Controllable abstractive summarization. *arXiv preprint arXiv:1711.05217* (2017).

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. *arXiv preprint arXiv:1707.02633* (2017).

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, Hong Kong, China, 70–79. https://doi.org/10.18653/v1/D19-5409

Anushka Gupta, Diksha Chugh, Rahul Katarya, et al. 2022. Automated news summarization using transformers. In *Sustainable Advanced Computing*. Springer, 249–259.

Prakhar Gupta, Jeffrey Bigham, Yulia Tsvetkov, and Amy Pavel. 2021. Controlling Dialogue Generation with Semantic Exemplars. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 3018–3029. https://doi.org/10.18653/v1/2021.naacl-main.240

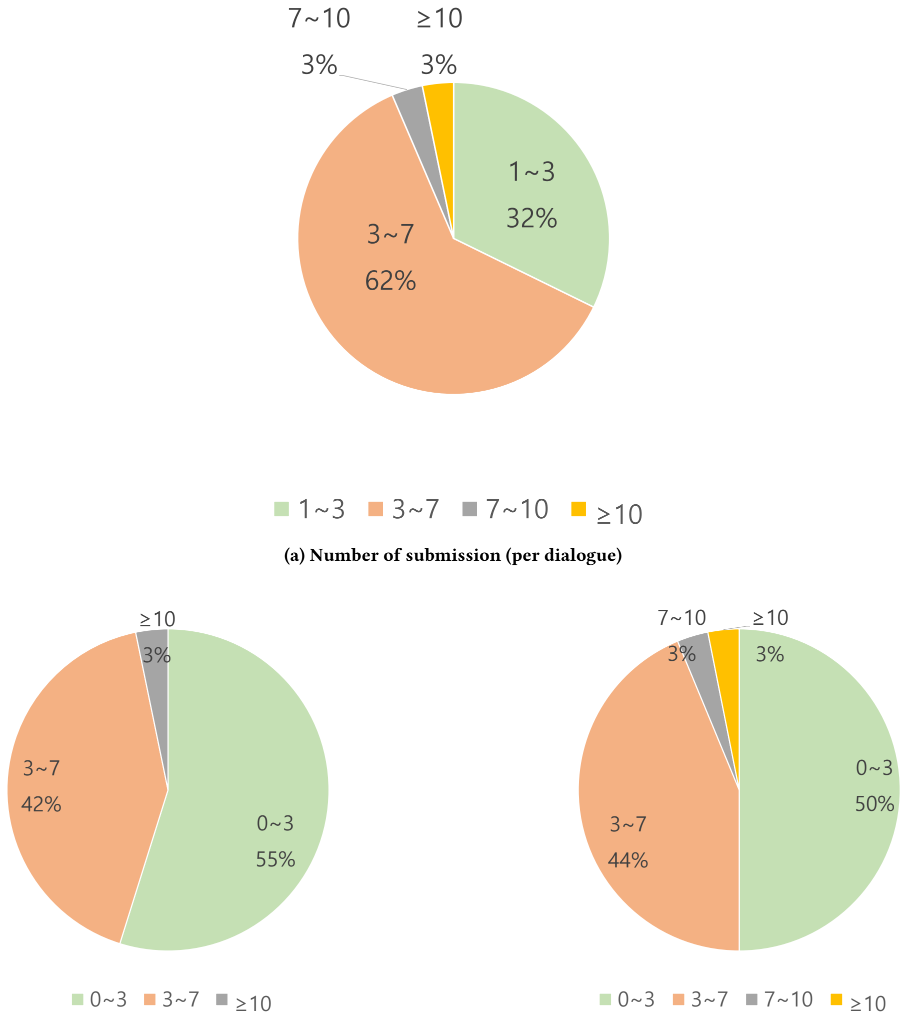
Prakhar Gupta, Jeffrey P Bigham, Yulia Tsvetkov, and Amy Pavel. 2020. Controlling dialogue generation with semantic exemplars. *arXiv preprint arXiv:2008.09075* (2020).

Matthew Guzdial, Nicholas Liao, Jonathan Chen, Shao-Yu Chen, Shukan Shah, Vishwa Shah, Joshua Reno, Gillian Smith, and Mark O Riedl. 2019. Friend, collaborator, student, manager: How design of an ai-driven game level editor affects creators. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J Cai. 2020. AI song contest: Human-AI co-creation in songwriting. *arXiv preprint arXiv:2010.05388* (2020).

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey. https://doi.org/10.48550/ARXIV.2104.14839

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, et al. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 955–964.



(a) Number of submission (per dialogue)

(b) Number of retouch WP (per dialogue)

(c) Number of retouch PP (per dialogue)

Figure 11: Summarization result satisfaction rate and user’s retouch count for automatically generated WP and PP tags.

Pegah Karimi, Mary Lou Maher, Nicholas Davis, and Kazjon Grace. 2019. Deep learning in a computational model for conceptual shifts in a co-creative design system. *arXiv preprint arXiv:1906.10188* (2019).

Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. 2019. May AI? Design ideation with cooperative contextual bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

Hui Liu, Huan Liu, Xin Wang, Wei Shao, Xiao Wang, Junzhao Du, Jonathan Liono, and Flora D Salim. 2020. SmartMeeting: An Novel Mobile Voice Meeting Minutes Generation and Analysis System. *Mobile Networks and Applications* 25, 2 (2020), 521–536.

Wei Liu, Huanqin Wu, Wenjing Mu, Zhen Li, Tao Chen, and Dan Nie. 2021. CO2Sum: Contrastive Learning for Factual-Consistent Abstractive Summarization. *CoRR* abs/2112.01147 (2021). arXiv:2112.01147 https://arxiv.org/abs/2112.01147

Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021a. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130* (2021).

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021b. Improving Factual Consistency of Abstractive Summarization via Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 6881–6894. https://doi.org/10.18653/v1/2021.acl-long.536

Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

Olabiyi Oluwatobi and Erik Mueller. 2020. DLGNet: A Transformer-based Model for Dialogue Response Generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, Online, 54–62. https://doi.org/10.18653/v1/2020.nlp4convai-1.7

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual* *meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.

Itsumi Saito, Kyosuke Nishida, Kosuke Nishida, Atsushi Otsuka, Hisako Asano, Junji Tomita, Hiroyuki Shindo, and Yuji Matsumoto. 2020. Length-controllable abstractive summarization by guiding with summary prototype. *arXiv preprint arXiv:2001.07331* (2020).

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053* (2019).

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188* (2022).

Ronnie Smith and Mauro Dragone. 2022. A Dialogue-Based Interface for Active Learning of Activities of Daily Living. In *27th International Conference on Intelligent User Interfaces*. 820–831.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021a. Plan-then-generate: Controlled data-to-text generation via planning. *arXiv preprint arXiv:2108.13740* (2021).

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021b. Plan-then-Generate: Controlled Data-to-Text Generation via Planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 895–909. https://doi.org/10.18653/v1/2021.findings-emnlp.76

Minhyang Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as social glue: uncovering the roles of deep generative AI during social music composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

Shunsuke Takeno, Masaaki Nagata, and Kazuhide Yamamoto. 2017. Controlling target features in neural machine translation via prefix constraints. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. 55–63.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

Jesse Vig, Wojciech Kryscinski, Karan Goel, and Nazneen Rajani. 2021. SummVis: Interactive Visual Analysis of Models, Data, and Evaluation for Text Summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 150–158. https://doi.org/10.18653/v1/2021.acl-demo.18

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2022. Question-aware transformer models for consumer health question summarization. *Journal of Biomedical Informatics* 128 (2022), 104040.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11765–11773.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing Factual Consistency of Abstractive Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 718–733. https://doi.org/10.18653/v1/2021.naacl-main.58

A DETAILED EXPERIMENTAL RESULTS

Tables 5 to 10 show the results on dialogue summarization evaluation on the WoW-food test based on T5 and BART.

Model	Beam Size	B-1	B-2	B-3	B-4	R-L	BERTScore	WP-Accuracy	PP-ROUGE-L
T5									
Baseline (w/o CML)	1	55.32(±0.4)	45.49(±0.6)	38.79(±0.8)	33.58(±0.9)	49.76(±0.7)	92.17(±0.1)	77.09(±1.1)	31.62(±0.4)
	3	57.30(±0.4)	47.70(±0.5)	41.03(±0.4)	35.76(±0.4)	51.70(±0.6)	92.55(±0.1)	81.16(±0.7)	32.45(±0.6)
	5	58.08(±0.3)	48.59(±0.5)	41.92(±0.5)	36.64(±0.5)	52.24(±0.7)	92.72(±0.1)	83.26(±0.7)	32.46(±0.5)
	10	58.77(±0.6)	49.41(±0.7)	42.75(±0.7)	37.46(±0.8)	52.76(±0.9)	92.86(±0.1)	85.59(±0.6)	32.60(±0.5)
	30	59.36(±0.4)	50.03(±0.5)	43.30(±0.5)	37.92(±0.5)	53.06(±0.6)	93.02(±0.0)	88.38(±0.7)	32.44(±0.4)
	50	59.70(±0.4)	50.33(±0.5)	43.53(±0.6)	38.07(±0.6)	53.24(±0.5)	93.06(±0.0)	89.33(±0.6)	32.68(±0.5)
	100	60.31(±0.5)	50.98(±0.6)	44.17(±0.6)	38.71(±0.6)	53.86(±0.9)	93.18(±0.1)	90.79(±0.6)	32.71(±0.5)
Only CML-sensitive fine-tuning	1	60.30(±1.0)	50.52(±1.0)	43.39(±0.9)	37.77(±0.9)	55.00(±0.7)	93.17(±0.2)	85.92(±1.0)	33.15(±0.4)
	3	62.06(±0.5)	52.50(±0.6)	45.44(±0.7)	39.74(±0.8)	56.36(±0.3)	93.46(±0.1)	88.69(±0.4)	34.01(±0.5)
	5	62.49(±0.4)	52.92(±0.4)	45.84(±0.4)	40.10(±0.4)	56.61(±0.3)	93.55(±0.1)	89.87(±0.6)	**34.02(±0.4)**
	10	62.88(±0.4)	53.39(±0.5)	46.31(±0.5)	40.59(±0.4)	57.18(±0.5)	93.65(±0.1)	91.10(±0.3)	34.00(±0.4)
	30	63.28(±0.6)	53.73(±0.5)	46.61(±0.5)	40.81(±0.5)	57.01(±0.5)	93.71(±0.0)	92.68(±0.3)	33.84(±0.3)
	50	63.32(±0.9)	53.80(±0.9)	46.70(±0.8)	40.94(±0.8)	57.11(±0.7)	93.73(±0.1)	93.53(±0.5)	33.87(±0.3)
	100	63.43(±0.6)	53.86(±0.7)	46.71(±0.8)	40.93(±0.9)	57.18(±1.0)	93.75(±0.1)	94.44(±0.4)	33.82(±0.4)
CML-sensitive pre-training + fine-tuning	1	68.16(±0.2)	59.23(±0.2)	52.30(±0.2)	46.53(±0.2)	61.36(±0.6)	94.44(±0.1)	91.93(±0.1)	33.60(±0.3)
	3	69.03(±0.3)	60.12(±0.3)	53.15(±0.3)	47.34(±0.3)	62.22(±0.4)	94.63(±0.1)	93.71(±0.3)	33.88(±0.3)
	5	69.26(±0.4)	60.33(±0.5)	53.33(±0.5)	47.49(±0.5)	62.19(±0.4)	94.66(±0.1)	94.58(±0.4)	33.81(±0.2)
	10	**69.57(±0.3)**	**60.61(±0.3)**	**53.56(±0.3)**	**47.69(±0.3)**	**62.27(±0.5)**	94.71(±0.1)	95.41(±0.3)	33.74(±0.3)
	30	69.42(±0.4)	60.42(±0.3)	53.34(±0.3)	47.46(±0.3)	62.22(±0.6)	94.71(±0.1)	96.41(±0.2)	33.53(±0.2)
	50	69.36(±0.4)	60.30(±0.3)	53.17(±0.3)	47.23(±0.3)	62.21(±0.6)	**94.72(±0.1)**	96.75(±0.2)	33.48(±0.2)
	100	69.26(±0.2)	60.22(±0.1)	53.09(±0.2)	47.16(±0.2)	62.06(±0.3)	94.71(±0.1)	**97.09(±0.1)**	33.36(±0.1)
BART									
Baseline (w/o CML)	1	55.19(±1.3)	44.90(±1.4)	38.04(±1.3)	32.74(±1.2)	49.19(±0.9)	92.05(±0.2)	74.00(±1.3)	31.38(±0.5)
	3	57.12(±1.5)	47.09(±1.7)	40.23(±1.7)	34.83(±1.5)	50.88(±1.1)	92.37(±0.3)	77.33(±1.2)	32.45(±0.5)
	5	57.74(±1.2)	47.76(±1.4)	40.94(±1.5)	35.53(±1.4)	51.38(±1.0)	92.54(±0.3)	78.70(±0.9)	32.81(±0.4)
	10	58.86(±1.1)	48.91(±1.3)	42.02(±1.5)	36.57(±1.4)	51.95(±1.0)	92.75(±0.2)	81.19(±0.9)	32.94(±0.4)
	30	60.12(±0.9)	50.17(±1.0)	43.14(±1.0)	37.59(±0.9)	53.00(±1.1)	92.97(±0.3)	84.29(±1.1)	33.32(±0.2)
	50	60.65(±0.7)	50.74(±0.9)	43.70(±1.0)	38.10(±0.9)	53.34(±1.2)	93.08(±0.2)	85.57(±1.2)	33.40(±0.3)
	100	61.35(±0.6)	51.43(±0.7)	44.36(±0.7)	38.68(±0.6)	53.63(±0.8)	93.18(±0.2)	87.16(±0.7)	33.36(±0.3)
Only CML-sensitive fine-tuning	1	58.77(±1.5)	48.72(±1.4)	41.62(±1.2)	35.97(±1.1)	53.29(±0.7)	92.76(±0.2)	80.84(2.2)	31.95(±0.9)
	3	60.73(±0.8)	50.75(±0.7)	43.52(±0.6)	37.74(±0.7)	54.71(±0.5)	93.08(±0.1)	83.50(±1.5)	32.66(±0.5)
	5	61.35(±0.6)	51.36(±0.4)	44.12(±0.3)	38.34(±0.2)	55.29(±0.8)	93.18(±0.1)	84.75(±1.5)	32.89(±0.6)
	10	61.99(±0.9)	51.95(±0.9)	44.65(±0.8)	38.81(±0.8)	55.35(±0.7)	93.26(±0.2)	86.71(±0.9)	33.23(±0.2)
	30	62.69(±0.6)	52.63(±0.6)	45.27(±0.6)	39.32(±0.7)	55.75(±0.6)	93.38(±0.1)	89.03(±0.8)	33.23(±0.2)
	50	63.07(±0.7)	53.04(±0.7)	45.70(±0.6)	39.75(±0.6)	56.12(±0.6)	93.47(±0.1)	90.02(±0.9)	33.37(±0.4)
	100	63.28(±1.1)	53.25(±1.1)	45.87(±1.1)	39.93(±1.1)	56.41(±0.7)	93.52(±0.1)	91.15(±1.0)	33.30(±0.3)
CML-sensitive pre-training + fine-tuning	1	67.67(±0.9)	58.61(±1.0)	51.58(±1.1)	45.74(±1.2)	62.03(±0.7)	94.36(±0.1)	90.06(±0.5)	32.93(±0.6)
	3	68.81(±0.5)	59.91(±0.7)	53.00(±0.8)	47.29(±1.0)	63.25(±0.7)	94.58(±0.1)	91.71(±0.5)	33.56(±0.3)
	5	69.41(±0.8)	60.45(±0.8)	53.50(±0.9)	47.80(±1.1)	63.58(±0.8)	94.68(±0.1)	92.79(±0.3)	33.57(±0.2)
	10	69.77(±0.7)	60.80(±0.9)	53.82(±1.0)	48.09(±1.2)	63.74(±1.0)	94.73(±0.2)	93.66(±0.4)	33.72(±0.3)
	30	70.12(±0.7)	61.12(±1.0)	54.09(±1.1)	48.30(±1.3)	63.82(±0.9)	94.78(±0.1)	95.01(±0.2)	**33.74(±0.5)**
	50	**70.22(±0.7)**	**61.23(±0.8)**	**54.23(±0.8)**	**48.44(±1.0)**	**63.83(±0.5)**	**94.80(±0.1)**	95.51(±0.3)	33.63(±0.4)
	100	70.05(±0.7)	61.09(±0.8)	54.07(±0.8)	48.22(±0.9)	63.64(±0.4)	94.79(±0.1)	**95.93(±0.2)**	**33.74(±0.4)**

Table 5: Experimental results of automatic metrics on the WoW-food test set. The reranking strategy is WP. For short, B and R refer to BLEU and ROUGE, respectively. Results are averaged over five random runs. The highest numbers are in bold.

Model	Beam Size	B-1	B-2	B-3	B-4	R-L	BERTScore	WP-Accuracy	PP-ROUGE-L
T5									
Baseline (w/o CML)	1	55.32(±0.4)	45.49(±0.6)	38.79(±0.8)	33.58(±0.9)	49.76(±0.7)	92.17(±0.1)	77.09(±1.1)	31.62(±0.4)
	3	56.80(±0.4)	47.20(±0.5)	40.64(±0.6)	35.48(±0.6)	51.55(±0.8)	92.43(±0.1)	77.79(±0.6)	34.00(±0.5)
	5	57.34(±0.6)	47.85(±0.6)	41.30(±0.6)	36.14(±0.7)	51.89(±0.5)	92.52(±0.1)	78.39(±1.0)	34.83(±0.4)
	10	57.63(±0.6)	48.25(±0.6)	41.75(±0.6)	36.61(±0.5)	52.01(±0.4)	92.58(±0.1)	78.83(±0.7)	35.87(±0.3)
	30	58.18(±0.4)	48.79(±0.5)	42.30(±0.5)	37.17(±0.5)	52.80(±0.5)	92.69(±0.1)	79.25(±1.0)	37.26(±0.3)
	50	58.10(±0.6)	48.69(±0.7)	42.17(±0.7)	37.01(±0.7)	52.78(±0.8)	92.68(±0.1)	79.27(±1.1)	37.90(±0.3)
	100	58.15(±0.4)	48.80(±0.4)	42.29(±0.5)	37.14(±0.4)	52.81(±0.5)	92.71(±0.1)	79.60(±0.7)	38.72(±0.4)
Only CML-sensitive fine-tuning	1	60.30(±1.0)	50.52(±1.0)	43.39(±0.9)	37.77(±0.9)	55.00(±0.7)	93.17(±0.2)	85.92(±1.0)	33.15(±0.4)
	3	61.66(±0.4)	52.18(±0.5)	45.19(±0.6)	39.56(±0.7)	56.49(±0.3)	93.40(±0.1)	86.56(±0.4)	35.15(±0.4)
	5	61.85(±0.4)	52.36(±0.4)	45.40(±0.5)	39.79(±0.5)	56.71(±0.5)	93.46(±0.1)	86.44(±0.5)	35.99(±0.4)
	10	61.94(±0.4)	52.48(±0.4)	45.50(±0.3)	39.87(±0.3)	56.78(±0.5)	93.51(±0.0)	86.51(±0.5)	36.98(±0.5)
	30	61.76(±0.7)	52.44(±0.7)	45.51(±0.5)	39.92(±0.5)	56.77(±0.5)	93.52(±0.1)	86.38(±0.8)	38.25(±0.4)
	50	61.66(±0.5)	52.32(±0.6)	45.40(±0.5)	39.87(±0.4)	56.88(±0.5)	93.54(±0.1)	86.33(±0.7)	38.79(±0.4)
	100	61.52(±0.5)	52.19(±0.5)	45.32(±0.5)	39.80(±0.3)	56.91(±0.7)	93.53(±0.1)	86.29(±0.7)	**39.46(±0.6)**
CML-sensitive pre-training + fine-tuning	1	68.16(±0.2)	59.23(±0.2)	52.30(±0.2)	46.53(±0.2)	61.36(±0.6)	94.44(±0.1)	91.93(±0.1)	33.60(±0.3)
	3	68.71(±0.4)	59.92(±0.3)	53.01(±0.4)	47.27(±0.4)	62.44(±0.3)	94.56(±0.0)	**92.24(±0.3)**	35.02(±0.2)
	5	**68.86(±0.3)**	**60.08(±0.3)**	53.19(±0.4)	47.45(±0.4)	62.83(±0.4)	94.62(±0.0)	92.12(±0.2)	35.62(±0.2)
	10	68.79(±0.2)	**60.08(±0.3)**	**53.21(±0.4)**	**47.48(±0.3)**	**63.34(±0.5)**	**94.65(±0.0)**	92.16(±0.6)	36.29(±0.2)
	30	68.45(±0.6)	59.64(±0.7)	52.78(±0.7)	47.09(±0.7)	63.16(±0.8)	94.60(±0.1)	91.94(±0.4)	37.26(±0.2)
	50	68.12(±0.8)	59.31(±0.9)	52.47(±0.9)	46.82(±0.9)	63.03(±0.8)	94.58(±0.1)	92.08(±0.2)	37.78(±0.2)
	100	67.97(±0.6)	59.23(±0.7)	52.45(±0.7)	46.81(±0.7)	63.25(±0.6)	94.60(±0.1)	91.69(±0.4)	38.32(±0.1)
BART									
Baseline (w/o CML)	1	55.19(±1.3)	44.90(±1.4)	38.04(±1.3)	32.74(±1.2)	49.19(±0.9)	92.05(±0.2)	74.00(±1.3)	31.38(±0.5)
	3	56.53(±1.3)	46.58(±1.5)	39.86(±1.5)	34.57(±1.4)	50.92(±0.8)	92.29(±0.2)	75.23(±1.1)	33.29(±0.5)
	5	56.71(±0.9)	46.85(±1.1)	40.16(±1.1)	34.91(±1.2)	51.23(±0.6)	92.36(±0.2)	75.31(±1.0)	34.22(±0.4)
	10	56.96(±1.0)	47.16(±1.0)	40.47(±1.0)	35.20(±1.1)	51.43(±0.8)	92.39(±0.2)	75.49(±1.0)	35.41(±0.1)
	30	57.45(±0.5)	47.80(±0.6)	41.18(±0.7)	35.96(±0.7)	52.21(±0.4)	92.48(±0.1)	76.02(±1.4)	37.20(±0.5)
	50	57.49(±0.4)	47.94(±0.6)	41.36(±0.6)	36.13(±0.6)	52.25(±0.4)	92.50(±0.1)	76.19(±1.4)	37.87(±0.4)
	100	57.50(±0.6)	48.12(±0.7)	41.62(±0.8)	36.47(±0.8)	52.71(±0.7)	92.54(±0.1)	76.42(±1.5)	**38.80(±0.4)**
Only CML-sensitive fine-tuning	1	58.77(±1.5)	48.72(±1.4)	41.62(±1.2)	35.97(±1.1)	53.29(±0.7)	92.76(±0.2)	80.84(2.2)	31.95(±0.9)
	3	60.41(±1.2)	50.52(±1.0)	43.35(±0.9)	37.63(±0.8)	54.67(±0.5)	93.05(±0.2)	82.12(2.0)	33.42(±0.4)
	5	60.53(±1.3)	50.70(±1.1)	43.62(±0.9)	37.95(±0.9)	55.12(±0.7)	93.09(±0.2)	82.33(2.3)	34.20(±0.4)
	10	60.65(±1.0)	50.96(±1.0)	43.94(±0.9)	38.33(±0.9)	55.41(±0.5)	93.11(±0.2)	82.71(±1.8)	35.17(±0.3)
	30	60.62(±1.2)	51.17(±1.0)	44.29(±0.8)	38.73(±0.8)	56.00(±0.5)	93.15(±0.1)	82.39(±1.2)	36.65(±0.5)
	50	60.68(±0.9)	51.22(±0.8)	44.36(±0.7)	38.85(±0.7)	56.35(±0.4)	93.16(±0.1)	82.44(±1.5)	37.31(±0.4)
	100	60.71(±1.1)	51.27(±0.9)	44.47(±0.8)	39.06(±0.7)	56.49(±0.6)	93.20(±0.1)	82.60(±1.3)	37.94(±0.4)
CML-sensitive pre-training + fine-tuning	1	67.67(±0.9)	58.61(±1.0)	51.58(±1.1)	45.74(±1.2)	62.03(±0.7)	94.36(±0.1)	90.06(±0.5)	32.93(±0.6)
	3	68.45(±0.2)	59.59(±0.5)	52.75(±0.6)	47.03(±0.7)	63.27(±0.5)	94.53(±0.1)	90.42(±0.5)	34.40(±0.3)
	5	**68.62(±0.7)**	**59.77(±0.6)**	**52.94(±0.5)**	**47.28(±0.7)**	63.54(±0.2)	**94.58(±0.1)**	90.51(±0.6)	34.99(±0.2)
	10	68.48(±0.8)	59.68(±0.7)	52.85(±0.6)	47.21(±0.7)	63.57(±0.4)	**94.58(±0.1)**	**90.74(±0.4)**	35.70(±0.3)
	30	68.23(±0.7)	59.44(±0.6)	52.65(±0.6)	47.05(±0.8)	63.53(±0.4)	94.52(±0.1)	90.42(±0.6)	36.99(±0.4)
	50	68.11(±0.8)	59.31(±0.7)	52.52(±0.6)	46.92(±0.5)	63.45(±0.5)	94.50(±0.1)	90.25(±0.5)	37.44(±0.3)
	100	68.05(±0.7)	59.39(±0.6)	52.69(±0.7)	47.15(±0.8)	**63.60(±0.5)**	94.48(±0.1)	90.01(±0.4)	38.10(±0.4)

Table 6: Experimental results of automatic metrics on the WoW-food test set. The reranking strategy is PP. For short, B and R refer to BLEU and ROUGE, respectively. Results are averaged over five random runs. The highest numbers are in bold.

Model	Beam Size	B-1	B-2	B-3	B-4	R-L	BERTScore	WP-Accuracy	PP-ROUGE-L
					T5				
Only CML-sensitive fine-tuning	1	58.85(±1.0)	49.04(±1.0)	41.95(±1.1)	36.33(±1.1)	53.16(±0.6)	92.92(±0.1)	84.55(±1.3)	32.18(±0.3)
	3	60.48(±0.4)	50.80(±0.6)	43.73(±0.6)	38.07(±0.6)	54.97(±0.3)	93.21(±0.0)	87.45(±0.3)	33.65(±0.2)
	5	61.49(±0.4)	51.88(±0.5)	44.78(±0.6)	39.08(±0.7)	55.58(±0.6)	93.35(±0.1)	88.82(±0.5)	34.32(±0.3)
	10	62.22(±0.4)	52.67(±0.4)	45.62(±0.5)	39.93(±0.6)	56.04(±0.6)	93.50(±0.1)	90.34(±0.5)	35.10(±0.5)
	30	63.23(±0.6)	53.63(±0.6)	46.53(±0.6)	40.74(±0.6)	56.64(±0.3)	93.68(±0.1)	92.23(±0.3)	35.90(±0.5)
	50	63.60(±0.6)	53.95(±0.6)	46.81(±0.5)	40.99(±0.5)	56.96(±0.1)	93.71(±0.1)	92.97(±0.2)	36.31(±0.5)
	100	63.68(±0.8)	54.04(±0.9)	46.90(±0.9)	41.09(±0.9)	57.01(±0.5)	93.74(±0.1)	93.62(±0.2)	**36.80(±0.4)**
CML-sensitive pre-training + fine-tuning	1	66.34(±0.5)	56.94(±0.7)	49.88(±0.8)	44.09(±0.8)	59.40(±0.8)	94.22(±0.1)	92.06(±0.5)	32.82(±0.3)
	3	67.35(±0.3)	58.11(±0.3)	51.09(±0.2)	45.27(±0.2)	60.73(±0.9)	94.41(±0.1)	93.68(±0.4)	34.19(±0.3)
	5	67.63(±0.4)	58.37(±0.4)	51.31(±0.3)	45.48(±0.3)	61.09(±0.6)	94.46(±0.1)	94.37(±0.5)	34.59(±0.4)
	10	67.81(±0.5)	58.61(±0.6)	51.59(±0.6)	45.77(±0.5)	61.42(±0.4)	94.50(±0.0)	95.14(±0.4)	35.13(±0.3)
	30	**68.10(±0.2)**	**59.03(±0.1)**	**52.01(±0.1)**	**46.15(±0.2)**	61.97(±0.2)	94.58(±0.0)	96.09(±0.1)	35.85(±0.4)
	50	67.95(±0.3)	58.85(±0.3)	51.84(±0.4)	46.00(±0.4)	**62.21(±0.3)**	**94.59(±0.0)**	96.32(±0.1)	36.28(±0.3)
	100	68.00(±0.3)	58.89(±0.5)	51.85(±0.6)	46.00(±0.7)	62.11(±0.6)	94.58(±0.1)	**96.80(±0.2)**	36.66(±0.1)
					BART				
Only CML-sensitive fine-tuning	1	58.81(±0.4)	48.70(±0.5)	41.62(±0.6)	35.98(±0.8)	52.87(±0.7)	92.76(±0.2)	81.43(±0.6)	31.75(±0.8)
	3	60.46(±0.9)	50.40(±1.0)	43.32(±1.0)	37.62(±0.9)	54.43(±0.8)	93.08(±0.2)	84.21(±1.5)	33.17(±0.6)
	5	61.29(±0.7)	51.27(±0.8)	44.15(±0.8)	38.39(±0.9)	55.06(±0.8)	93.22(±0.2)	85.49(±1.3)	33.75(±0.6)
	10	61.89(±0.6)	51.86(±0.6)	44.69(±0.6)	38.89(±0.8)	55.63(±0.6)	93.36(±0.1)	86.93(±1.0)	34.53(±0.6)
	30	62.73(±1.0)	52.84(±1.1)	45.65(±1.1)	39.86(±1.1)	56.58(±1.1)	93.50(±0.1)	89.38(±1.2)	35.32(±0.5)
	50	63.00(±0.8)	53.08(±0.7)	45.88(±0.6)	40.07(±0.7)	56.96(±1.0)	93.55(±0.1)	90.11(±1.1)	35.76(±0.6)
	100	63.60(±0.6)	53.77(±0.7)	46.58(±0.7)	40.75(±0.8)	57.31(±0.4)	93.65(±0.1)	91.16(±0.8)	**36.33(±0.7)**
CML-sensitive pre-training + fine-tuning	1	65.94(±0.7)	56.58(±0.8)	49.50(±0.7)	43.77(±0.7)	59.88(±0.6)	94.07(±0.1)	90.32(±0.6)	32.22(±0.5)
	3	67.26(±0.4)	58.04(±0.5)	50.99(±0.7)	45.24(±0.7)	61.56(±0.9)	94.31(±0.1)	91.98(±0.6)	33.22(±0.4)
	5	67.61(±0.6)	58.44(±0.7)	51.41(±0.8)	45.67(±0.9)	61.85(±1.2)	94.41(±0.1)	92.97(±0.5)	33.66(±0.6)
	10	68.13(±0.9)	58.91(±0.8)	51.82(±0.7)	45.99(±0.7)	62.05(±0.9)	94.47(±0.1)	94.00(±0.6)	34.25(±0.5)
	30	68.48(±0.3)	59.26(±0.3)	52.21(±0.2)	46.39(±0.2)	62.52(±0.5)	94.54(±0.1)	95.23(±0.4)	35.18(±0.6)
	50	68.54(±0.3)	59.27(±0.4)	52.17(±0.4)	46.35(±0.4)	62.53(±0.4)	94.54(±0.1)	95.62(±0.3)	35.62(±0.4)
	100	**68.91(±0.3)**	**59.65(±0.4)**	**52.51(±0.4)**	**46.67(±0.5)**	**62.73(±0.4)**	**94.58(±0.1)**	**96.13(±0.2)**	36.06(±0.4)

Table 7: Experimental results of automatic metrics on the WoW-food test set with WP tags only. The reranking strategy is SUM. For short, B and R refer to BLEU and ROUGE, respectively. Results are averaged over five random runs. The highest numbers are in bold.

Model	Beam Size	B-1	B-2	B-3	B-4	R-L	BERTScore	WP-Accuracy	PP-ROUGE-L
					T5				
Only CML-sensitive fine-tuning	1	58.85(±1.0)	49.04(±1.0)	41.95(±1.1)	36.33(±1.1)	53.16(±0.6)	92.92(±0.1)	84.55(±1.3)	32.18(±0.3)
	3	60.29(±0.3)	50.56(±0.4)	43.49(±0.5)	37.83(±0.5)	54.79(±0.5)	93.17(±0.1)	87.51(±0.3)	32.88(±0.1)
	5	61.02(±0.5)	51.38(±0.5)	44.28(±0.5)	38.58(±0.6)	55.20(±0.4)	93.28(±0.1)	88.92(±0.4)	33.14(±0.2)
	10	61.83(±0.4)	52.20(±0.4)	45.10(±0.5)	39.39(±0.5)	55.56(±0.5)	93.41(±0.1)	90.58(±0.5)	**33.31(±0.4)**
	30	62.33(±0.6)	52.68(±0.5)	45.57(±0.6)	39.84(±0.6)	55.91(±0.3)	93.55(±0.0)	92.60(±0.2)	33.28(±0.3)
	50	62.52(±0.8)	52.79(±0.7)	45.60(±0.7)	39.81(±0.7)	56.00(±0.6)	93.55(±0.1)	93.44(±0.2)	33.28(±0.5)
	100	62.79(±0.4)	53.13(±0.3)	45.97(±0.3)	40.17(±0.3)	56.17(±0.3)	93.58(±0.0)	94.15(±0.3)	32.99(±0.4)
CML-sensitive pre-training + fine-tuning	1	66.34(±0.5)	56.94(±0.7)	49.88(±0.8)	44.09(±0.8)	59.40(±0.8)	94.22(±0.1)	92.06(±0.5)	32.82(±0.3)
	3	67.22(±0.4)	57.89(±0.3)	50.79(±0.3)	44.95(±0.3)	60.23(±0.9)	94.38(±0.1)	93.77(±0.5)	33.28(±0.3)
	5	67.42(±0.3)	58.08(±0.3)	50.93(±0.3)	45.08(±0.3)	60.52(±0.5)	94.43(±0.1)	94.48(±0.4)	33.25(±0.2)
	10	**67.63(±0.4)**	**58.28(±0.5)**	**51.10(±0.4)**	**45.20(±0.3)**	60.61(±0.5)	94.46(±0.1)	95.29(±0.3)	33.23(±0.3)
	30	**67.63(±0.2)**	**58.28(±0.3)**	51.08(±0.4)	45.16(±0.3)	**60.75(±0.6)**	**94.47(±0.1)**	96.28(±0.1)	33.09(±0.2)
	50	67.38(±0.2)	58.06(±0.3)	50.87(±0.4)	44.96(±0.3)	60.53(±0.8)	94.44(±0.1)	96.62(±0.2)	33.01(±0.3)
	100	67.26(±0.4)	57.89(±0.5)	50.71(±0.6)	44.79(±0.6)	60.25(±1.1)	94.43(±0.1)	**97.11(±0.1)**	32.80(±0.3)
					BART				
Only CML-sensitive fine-tuning	1	58.81(±0.4)	48.70(±0.5)	41.62(±0.6)	35.98(±0.8)	52.87(±0.7)	92.76(±0.2)	81.43(±0.6)	31.75(±0.8)
	3	60.52(±0.9)	50.45(±1.0)	43.36(±1.0)	37.62(±0.9)	54.23(±0.9)	93.08(±0.2)	84.25(±1.5)	32.70(±0.4)
	5	61.22(±0.5)	51.15(±0.7)	44.00(±0.8)	38.22(±0.9)	54.80(±0.9)	93.18(±0.2)	85.57(±1.2)	32.91(±0.7)
	10	61.76(±0.5)	51.74(±0.5)	44.59(±0.7)	38.80(±0.8)	55.30(±0.8)	93.31(±0.1)	87.09(±1.0)	33.35(±0.6)
	30	62.51(±0.7)	52.56(±0.9)	45.39(±0.8)	39.60(±0.8)	55.73(±0.9)	93.44(±0.1)	89.63(±1.1)	33.26(±0.5)
	50	62.93(±0.4)	52.89(±0.6)	45.65(±0.8)	39.79(±0.8)	56.22(±0.8)	93.51(±0.0)	90.49(±0.9)	33.20(±0.6)
	100	63.36(±0.3)	53.36(±0.5)	46.08(±0.7)	40.14(±0.8)	56.44(±0.8)	93.60(±0.1)	91.57(±0.7)	**33.49(±0.9)**
CML-sensitive pre-training + fine-tuning	1	65.94(±0.7)	56.58(±0.8)	49.50(±0.7)	43.77(±0.7)	59.88(±0.6)	94.07(±0.1)	90.32(±0.6)	32.22(±0.5)
	3	67.14(±0.4)	57.88(±0.4)	50.79(±0.4)	45.02(±0.5)	61.30(±0.9)	94.31(±0.1)	92.03(±0.5)	32.71(±0.3)
	5	67.57(±0.7)	58.34(±0.7)	51.26(±0.7)	45.49(±0.8)	61.63(±1.0)	94.39(±0.1)	93.00(±0.5)	32.96(±0.5)
	10	68.02(±1.0)	58.74(±1.0)	51.59(±0.9)	45.75(±1.0)	61.75(±0.9)	94.47(±0.1)	94.06(±0.6)	33.10(±0.6)
	30	68.39(±0.4)	59.04(±0.4)	**51.88(±0.4)**	**46.03(±0.5)**	**62.03(±0.6)**	94.52(±0.1)	95.34(±0.4)	33.19(±0.6)
	50	68.45(±0.5)	59.06(±0.5)	51.87(±0.6)	46.00(±0.6)	61.83(±0.5)	94.53(±0.1)	95.81(±0.2)	33.12(±0.5)
	100	**68.54(±0.7)**	**59.06(±0.7)**	51.79(±0.6)	45.88(±0.5)	61.88(±0.4)	**94.54(±0.1)**	**96.33(±0.2)**	33.21(±0.6)

Table 8: Experimental results of automatic metrics on the WoW-food test set with WP tags only. The reranking strategy is WP. For short, B and R refer to BLEU and ROUGE, respectively. Results are averaged over five random runs. The highest numbers are in bold.

Model	Beam Size	B-1	B-2	B-3	B-4	R-L	BERTScore	WP-Accuracy	PP-ROUGE-L
T5									
Only CML-sensitive fine-tuning	1	57.40(±0.8)	47.58(±1.0)	40.77(±1.1)	35.37(±1.0)	51.88(±0.6)	92.62(±0.1)	79.08(±1.0)	32.52(±0.4)
	3	59.19(±0.5)	49.67(±0.4)	42.91(±0.3)	37.50(±0.3)	53.45(±0.4)	92.93(±0.1)	82.64(±1.0)	34.02(±0.3)
	5	59.81(±0.3)	50.33(±0.3)	43.55(±0.3)	38.07(±0.4)	54.22(±0.2)	93.03(±0.0)	84.11(±0.5)	34.62(±0.4)
	10	60.21(±0.5)	50.72(±0.7)	43.91(±0.8)	38.47(±0.9)	54.34(±0.5)	93.12(±0.1)	86.14(±0.4)	35.17(±0.2)
	30	61.26(±0.4)	51.70(±0.5)	44.82(±0.6)	39.32(±0.6)	55.21(±0.3)	93.31(±0.1)	88.60(±0.6)	35.88(±0.3)
	50	61.41(±0.5)	51.87(±0.5)	44.98(±0.5)	39.38(±0.5)	55.38(±0.5)	93.36(±0.1)	89.54(±0.2)	36.32(±0.4)
	100	61.86(±0.3)	52.36(±0.3)	45.46(±0.3)	39.87(±0.3)	55.37(±0.4)	93.43(±0.1)	90.79(±0.4)	36.95(±0.6)
CML-sensitive pre-training + fine-tuning	1	62.87(±0.4)	53.54(±0.5)	46.54(±0.5)	40.84(±0.6)	57.75(±0.5)	93.59(±0.1)	84.65(±0.6)	34.10(±0.5)
	3	64.47(±0.7)	55.17(±0.7)	48.19(±0.6)	42.45(±0.6)	59.18(±0.7)	93.89(±0.1)	87.41(±0.7)	35.29(±0.2)
	5	64.83(±0.4)	55.64(±0.3)	48.69(±0.2)	42.97(±0.2)	59.57(±0.4)	93.96(±0.0)	88.50(±0.5)	35.70(±0.3)
	10	65.15(±0.3)	56.07(±0.3)	49.18(±0.2)	43.49(±0.2)	60.00(±0.4)	94.06(±0.0)	90.03(±0.3)	36.17(±0.2)
	30	65.84(±0.6)	56.79(±0.7)	49.92(±0.6)	44.24(±0.5)	60.64(±0.5)	94.20(±0.1)	91.69(±0.3)	37.00(±0.2)
	50	65.95(±0.8)	56.92(±0.9)	50.04(±0.8)	44.33(±0.7)	61.01(±0.5)	94.22(±0.1)	92.35(±0.3)	37.30(±0.2)
	100	**66.04(±0.7)**	**57.09(±0.8)**	**50.25(±0.7)**	**44.55(±0.6)**	**61.30(±0.3)**	94.26(±0.1)	**93.32(±0.5)**	**37.60(±0.3)**
BART									
Only CML-sensitive fine-tuning	1	57.23(±1.1)	47.22(±1.2)	40.30(±1.3)	34.97(±1.4)	51.59(±1.0)	92.49(±0.2)	77.80(±1.3)	32.52(±1.0)
	3	59.25(±1.1)	49.29(±1.3)	42.32(±1.4)	36.91(±1.5)	53.11(±1.2)	92.84(±0.2)	80.92(±1.3)	33.82(±0.8)
	5	59.69(±1.6)	49.75(±1.8)	42.78(±1.8)	37.36(±1.8)	53.35(±1.2)	92.91(±0.3)	82.17(±1.7)	34.36(±1.0)
	10	60.68(±1.0)	50.84(±1.4)	43.91(±1.5)	38.51(±1.6)	54.34(±1.3)	93.08(±0.3)	83.89(±1.5)	35.17(±0.8)
	30	61.95(±1.1)	52.09(±1.4)	45.02(±1.5)	39.45(±1.5)	55.55(±1.4)	93.32(±0.3)	86.63(±1.4)	36.16(±0.6)
	50	62.38(±1.1)	52.54(±1.4)	45.47(±1.6)	39.78(±1.6)	55.69(±1.3)	93.36(±0.2)	87.63(±1.3)	36.46(±0.6)
	100	62.63(±0.8)	52.82(±0.9)	45.75(±1.0)	40.08(±1.1)	56.02(±1.1)	93.47(±0.2)	88.95(±0.8)	36.87(±0.7)
CML-sensitive pre-training + fine-tuning	1	62.78(±0.7)	53.62(±0.7)	46.79(±0.8)	41.28(±0.9)	57.97(±0.7)	93.49(±0.1)	83.56(±1.2)	34.15(±0.5)
	3	64.20(±0.7)	55.13(±1.0)	48.31(±1.2)	42.79(±1.4)	59.48(±0.9)	93.80(±0.1)	86.05(±1.0)	35.24(±0.4)
	5	64.66(±0.6)	55.58(±0.8)	48.73(±1.0)	43.20(±1.2)	59.70(±0.8)	93.88(±0.1)	86.79(±0.3)	35.71(±0.6)
	10	65.02(±0.8)	55.96(±1.0)	49.04(±1.1)	43.47(±1.2)	59.99(±1.0)	93.95(±0.2)	88.06(±0.6)	36.44(±0.4)
	30	65.72(±0.8)	56.63(±0.8)	49.71(±0.9)	44.09(±0.9)	60.36(±0.5)	94.04(±0.1)	89.97(±0.4)	37.11(±0.4)
	50	65.97(±0.9)	56.89(±0.9)	49.96(±1.0)	44.29(±0.9)	60.55(±0.7)	94.11(±0.1)	90.67(±0.1)	37.49(±0.4)
	100	**66.43(±0.9)**	**57.25(±1.0)**	**50.22(±1.1)**	**44.49(±1.0)**	**60.74(±0.7)**	**94.16(±0.2)**	**91.63(±0.3)**	**37.80(±0.5)**

Table 9: Experimental results of automatic metrics on the WoW-food test set with PP tags only. The reranking strategy is SUM. For short, B and R refer to BLEU and ROUGE, respectively. Results are averaged over five random runs. The highest numbers are in bold.

Model	Beam Size	B-1	B-2	B-3	B-4	R-L	BERTScore	WP-Accuracy	PP-ROUGE-L
T5									
Only CML-sensitive fine-tuning	1	57.40(±0.8)	47.58(±1.0)	40.77(±1.1)	35.37(±1.0)	51.88(±0.6)	92.62(±0.1)	79.08(±1.0)	32.52(±0.4)
	3	58.55(±0.5)	49.00(±0.5)	42.26(±0.5)	36.92(±0.5)	53.48(±0.4)	92.84(±0.1)	80.46(±1.2)	34.56(±0.4)
	5	58.77(±0.4)	49.25(±0.4)	42.56(±0.3)	37.20(±0.3)	53.94(±0.5)	92.86(±0.1)	80.70(±0.8)	35.42(±0.5)
	10	58.91(±0.3)	49.39(±0.3)	42.68(±0.3)	37.37(±0.5)	54.05(±0.6)	92.87(±0.1)	80.67(±0.6)	36.45(±0.3)
	30	59.08(±0.3)	49.57(±0.4)	42.90(±0.5)	37.63(±0.5)	54.51(±0.5)	92.91(±0.1)	80.94(±0.6)	37.76(±0.4)
	50	59.19(±0.4)	49.66(±0.5)	42.97(±0.4)	37.66(±0.5)	54.52(±0.3)	92.93(±0.1)	81.20(±0.7)	38.40(±0.3)
	100	59.20(±0.5)	49.75(±0.6)	43.08(±0.5)	37.77(±0.5)	54.30(±0.6)	92.94(±0.1)	81.55(±0.6)	39.17(±0.4)
CML-sensitive pre-training + fine-tuning	1	62.87(±0.4)	53.54(±0.5)	46.54(±0.5)	40.84(±0.6)	57.75(±0.5)	93.59(±0.1)	84.65(±0.6)	34.10(±0.5)
	3	63.80(±0.7)	54.55(±0.7)	47.60(±0.8)	41.88(±0.8)	58.90(±0.6)	93.77(±0.1)	**85.61(±1.4)**	35.72(±0.2)
	5	63.78(±0.7)	54.65(±0.6)	47.77(±0.5)	42.11(±0.4)	59.08(±0.3)	93.76(±0.1)	85.29(±1.2)	36.43(±0.3)
	10	**63.84(±0.3)**	**54.75(±0.3)**	**47.88(±0.4)**	**42.25(±0.5)**	59.48(±0.4)	**93.81(±0.1)**	85.34(±0.4)	37.31(±0.3)
	30	63.63(±0.8)	54.52(±0.8)	47.69(±0.7)	42.10(±0.6)	**59.60(±0.4)**	93.79(±0.1)	84.88(±0.8)	38.81(±0.2)
	50	63.20(±0.7)	54.12(±0.7)	47.31(±0.7)	41.72(±0.7)	59.25(±0.5)	93.72(±0.1)	84.66(±0.7)	39.31(±0.2)
	100	63.01(±0.5)	53.98(±0.5)	47.20(±0.5)	41.68(±0.5)	59.50(±0.4)	93.74(±0.1)	84.69(±0.8)	**39.91(±0.3)**
BART									
Only CML-sensitive fine-tuning	1	57.23(±1.1)	47.22(±1.2)	40.30(±1.3)	34.97(±1.4)	51.59(±1.0)	92.49(±0.2)	77.80(±1.3)	32.52(±1.0)
	3	58.56(±1.1)	48.65(±1.3)	41.75(±1.4)	36.40(±1.5)	52.81(±1.0)	92.73(±0.3)	79.07(±1.8)	34.20(±0.9)
	5	58.83(±1.5)	48.96(±1.8)	42.09(±1.7)	36.76(±1.7)	53.06(±1.3)	92.77(±0.3)	79.43(±2.0)	34.97(±1.1)
	10	59.30(±1.2)	49.50(±1.6)	42.67(±1.6)	37.39(±1.6)	53.57(±1.2)	92.86(±0.3)	79.53(±1.7)	36.13(±0.7)
	30	59.46(±1.0)	49.74(±1.4)	42.92(±1.4)	37.60(±1.4)	54.33(±1.0)	92.91(±0.2)	79.87(±1.1)	37.71(±0.8)
	50	59.21(±0.9)	49.62(±1.2)	42.85(±1.3)	37.54(±1.4)	54.17(±1.1)	92.86(±0.2)	79.57(±1.3)	38.33(±0.6)
	100	59.09(±0.9)	49.44(±1.2)	42.69(±1.4)	37.43(±1.4)	54.20(±0.9)	92.85(±0.2)	79.67(±1.4)	39.11(±0.7)
CML-sensitive pre-training + fine-tuning	1	62.78(±0.7)	53.62(±0.7)	46.79(±0.8)	41.28(±0.9)	57.97(±0.7)	93.49(±0.1)	83.56(±1.2)	34.15(±0.5)
	3	63.38(±0.8)	54.33(±1.0)	47.59(±1.1)	42.17(±1.3)	59.19(±0.8)	93.66(±0.2)	83.92(±1.3)	35.69(±0.4)
	5	**63.43(±0.6)**	**54.43(±0.8)**	**47.70(±0.8)**	**42.27(±1.0)**	**59.32(±0.6)**	**93.67(±0.1)**	83.69(±0.7)	36.32(±0.5)
	10	63.30(±0.7)	54.34(±0.9)	47.60(±1.0)	42.16(±0.9)	59.29(±1.0)	93.66(±0.2)	**83.89(±0.9)**	37.36(±0.4)
	30	63.00(±0.2)	54.04(±0.4)	47.36(±0.6)	41.95(±0.6)	59.27(±0.5)	93.59(±0.1)	83.54(±0.7)	38.68(±0.5)
	50	63.04(±0.6)	54.11(±0.6)	47.44(±0.7)	42.03(±0.6)	59.30(±0.4)	93.63(±0.2)	83.53(±1.0)	39.37(±0.4)
	100	62.86(±0.8)	54.03(±0.7)	47.46(±0.9)	42.13(±1.0)	59.24(±0.6)	93.58(±0.2)	82.99(±0.6)	**39.98(±0.4)**

Table 10: Experimental results of automatic metrics on the WoW-food test set with PP tags only. The reranking strategy is PP. For short, B and R refer to BLEU and ROUGE, respectivelyt. Results are averaged over five random runs. The highest numbers are in bold.