STRUDEL: Structured Dialogue Summarization for Dialogue Comprehension

Borui Wang[1] **Chengcheng Feng**[1] **Arjun Nair**[1] **Madelyn Mao**[1] **Jai Desai**[1]
Asli Celikyilmaz[2] **Haoran Li**[2] **Yashar Mehdad**[2] **Dragomir Radev**[1]
[1]Yale University [2]Meta AI
{borui.wang, dragomir.radev}@yale.edu {aslic, aimeeli, mehdad}@fb.com

Abstract

Abstractive dialogue summarization has long been viewed as an important standalone task in natural language processing, but no previous work has explored the possibility of whether abstractive dialogue summarization can also be used as a means to boost an NLP system's performance on other important dialogue comprehension tasks. In this paper, we propose a novel type of dialogue summarization task - STRUctured DiaLoguE Summarization (STRUDEL) - that can help pre-trained language models to better understand dialogues and improve their performance on important dialogue comprehension tasks. In contrast to the holistic approach taken by the traditional free-form abstractive summarization task for dialogues, STRUDEL aims to decompose and imitate the hierarchical, systematic and structured mental process that we human beings usually go through when understanding and analyzing dialogues, and thus has the advantage of being more focused, specific and instructive for dialogue comprehension models to learn from. We further introduce a new STRUDEL dialogue comprehension modeling framework that integrates STRUDEL into a dialogue reasoning module over transformer encoder language models to improve their dialogue comprehension ability. In our empirical experiments on two important downstream dialogue comprehension tasks - dialogue question answering and dialogue response prediction - we demonstrate that our STRUDEL dialogue comprehension models can significantly improve the dialogue comprehension performance of transformer encoder language models.

1 Introduction

In natural language processing, abstractive dialogue summarization (Feng et al., 2021) has long been viewed as an important standalone task, but no previous work has explored the possibility of whether abstractive dialogue summarization can also be used as a means to boost an NLP system's performance on other important dialogue comprehension tasks. When performing language understanding, a very natural and effective first step that human beings usually take in their mental process is to try to summarize the main content of a piece of text, usually from multiple perspectives each focusing a different aspect of the text. This is especially true when human readers or speakers are trying to understand a dialogue or a conversation, which involve multi-turn exchange of information following a general theme, topic or storyline. Therefore, we would like to ask the following question - can the task of abstractive dialogue summarization also help NLP models to learn to perform better dialogue comprehension?

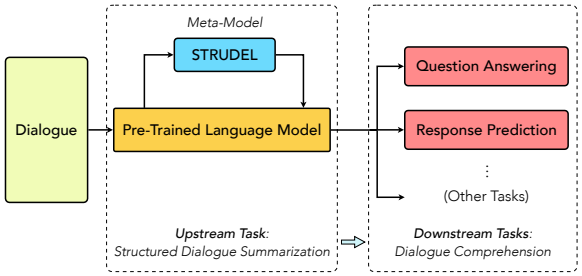


Figure 1: STRUDEL as a meta-model on top of pre-trained language models for dialogue comprehension.

In this paper, we propose a novel type of dialogue summarization task - STRUctured DiaLoguE Summarization (STRUDEL[1]) - that can help pre-trained language models to better understand dialogues and improve their performance on important dialogue comprehension tasks. In contrast to the holistic approach taken by the traditional free-form abstractive summarization task for dialogues, STRUDEL aims to decompose and imitate the hierarchical, systematic and structured mental process

[1]The name STRUDEL comes from a type of layered pastry with fillings called *strudel*, which, like our proposed task of structured dialogue summarization, is also structured.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4949 - 4958
December 7-11, 2022

that we human beings usually go through when understanding and analyzing dialogues. Then we further introduce a new dialogue comprehension model that integrates STRUDEL into a dialogue reasoning module over transformer encoder language models. Our empirical experiment results shows that STRUDEL is indeed very effective in providing transformer language models with better support for reasoning and inference over challenging downstream dialogue comprehension tasks such as dialogue question answering and response prediction and improving their performance.

2 Background and Related Work

2.1 Abstractive Summarization

Abstractive summarization aims to generate a concise summary of a text by producing a paraphrasing of the main contents using different vocabulary, rather than simply extracting the important sentences, which is referred to as extractive summarization. A popular approach to produce abstractive summaries of long documents is via neural abstractive summarization by using a singular extractive step to condition the transformer language model before generating a summary (Zhang and Zhao, 2021). Some other methods also take the structure of the dialogues into consideration when generating a single free-form abstractive summarization. For example, Wu et al. (2021) presented BASS, a novel framework for Boosting Abstractive Summarization based on a unified Semantic graph and a graph-based encoder-decoder model to improve summary generation process by leveraging the graph structure. Villmow et al. (2021) improved source code summarization tasks using self-attention with relative position representations to consider structural relationships between nodes which can encode movements between any pair of nodes in the tree.

Abstractive summarization has also been applied to solve NLP-related tasks such as text classification, news summarization, and headline generation. Furthermore, the generation of summaries can be integrated into these systems as an intermediate stage to reduce the length of documents. Mahalakshmi and Fatima (2022) presented a new text summarization model to retrieve information with deep learning methods. Du and Gao (2021) migrated the large-scale generic summarization datasets into query-focused datasets and proposed a model called SQAS, which can extract the reasoning information by understanding the source document via the question-answering model.

2.2 Dialogue Comprehension and Understanding

Abstractive dialogue summarization, the task of summarizing multi-turn conversations between different speakers (Feng et al., 2021), presents many additional challenges when compared to a narrative setting. For example, when attempting coreference resolution, text summarization models will often misattribute the actions, intentions, or statements of one speaker to another and fail to accurately model topic drift and diverse interactions across utterances (Feng et al., 2020a). Thus, it is especially important to specifically develop models that are capable of reasoning in a multi-turn dialogue setting for abstractive dialogue summarization.

There have been a number of advances in multi-turn dialogue comprehension and reasoning in recent years. Liu et al. (2020) showed that explicitly modeling speaker information for each token helped the summarization model resolve coreference errors. Ouyang et al. (2020) showed that separating the dialogue context into elementary discourse units (EDUs) and then modeling the relationship between those EDUs as a graph helped the model better understand the innate structure of the dialogue. Commonsense knowledge injection has been shown to improve the performance of dialogue summarization models (Feng et al., 2020b). Neural-retrieval-in-the-loop architectures have been shown to reduce hallucination in models (Shuster et al., 2021). Additionally, contrastive learning, which uses negative samples to show the model examples of what not to output, have seen increasing use across the field of abstractive summarization (Liu and Liu, 2021). For example, utterance inversion can help the model learn an implicit understanding of the temporal relationship between utterances.

3 Structured Dialogue Summarization

3.1 Definition of Structured Dialogue Summarization

We define *Structured Dialogue Summarization* (STRUDEL) as the task of generating a systematic and abstactive multi-entry dialogue summarization organized in a structured form that represents a comprehensive multi-aspect understanding and interpretation of a dialogue's content.

A complete STRUDEL summarization of a dialogue[2] contains a set of 16 STRUDEL entries, which are each defined as follows:

(a) **Name**$_{S_1}$ - the name of the first speaker of the dialogue.

(b) **Name**$_{S_2}$ - the name of the second speaker of the dialogue.

(c) **Role/Identity**$_{S_1}$ - the role or identity of the first speaker of the dialogue.

(d) **Role/Identity**$_{S_2}$ - the role or identity of the second speaker of the dialogue.

(e) **Relationship** - the relationship between the two speakers of the dialogue.

(f) **Time** - the time that the dialogue takes place.

(g) **Location**$_{S_1}$ - the physical location of the first speaker when the dialogue takes place.

(h) **Location**$_{S_2}$ - the physical location of the second speaker when the dialogue takes place.

(i) **Purpose/Theme** - the main purpose or theme for which the dialogue is made between the two speakers.

(j) **Task/Intention**$_{S_1}$ - the main task or intention that the first speaker would like to achieve in the dialogue.

(k) **Task/Intention**$_{S_2}$ - the main task or intention that the second speaker would like to achieve in the dialogue.

(l) **Problem/Disagreement**$_1$ - the most important problem or disagreement that the two speakers need to solve in the dialogue.

(m) **Solution**$_1$ - the solution that the two speakers reach for the most important problem or disagreement in the dialogue.

(n) **Problem/Disagreement**$_2$ - the second most important problem or disagreement that the two speakers need to solve in the dialogue.

(o) **Solution**$_2$ - the solution that the two speakers reach for the second most important problem or disagreement in the dialogue.

(p) **Conclusion/Agreement** - the final conclusion or agreement that the two speakers reach in the dialogue.

In an actual STRUDEL summarization of a dialogue, the content of each of the above 16 STRUDEL entries will either be a short text abstractively summarizing a specific aspect of the dialogue as indicated by that STRUDEL entry's definition, or be `'N/A'` indicating that the entry can't be inferred from or is not mentioned in the current dialogue.

3.2 Example of Structured Dialogue Summarization

Here we use a concrete example to demonstrate structured dialogue summarization of a dialogue. Figure 2 shows an example dialogue from the DREAM dataset (Sun et al., 2019). For this dialogue, its structured dialogue summarization is:

Name$_{S_1}$: *"N/A"*
Name$_{S_2}$: *"Bill."*
Role/Identity$_{S_1}$: *"Mother."*
Role/Identity$_{S_2}$: *"Father."*
Relationship: *"Wife and husband."*
Time: *"N/A"*.
Location$_{S_1}$: *"N/A"*
Location$_{S_2}$: *"N/A"*
Purpose/Theme: *"Go to the cinema this weekend."*
Task/Intention$_{S_1}$: *"Pick a movie to watch."*
Task/Intention$_{S_2}$: *"Pick a movie to watch."*
Problem/Disagreement$_1$: *"It's boring for Bill to watch the film Happy Potter and the Sorcerer's Stone."*
Solution$_1$: *"Bill will watch another film called the Most Wanted."*
Problem/Disagreement$_2$: *"N/A"*
Solution$_2$: *"N/A"*
Conclusion/Agreement: *"Go to the cinema and come home together, but watch different films."*

This same example also appears in the DIALOGSUM dataset (Chen et al., 2021), which is a dataset for traditional abstractive dialogue summarization. In contrast, this dialogue's traditional abstactive

[2]In this paper we focus on the structured dialogue summarization of two-speaker dialogues, which are the most commonly seen type of dialogues in dialogue datasets and real applications. We leave the extension of STRUDEL to multi-speaker dialogues to future work (see Section 8).

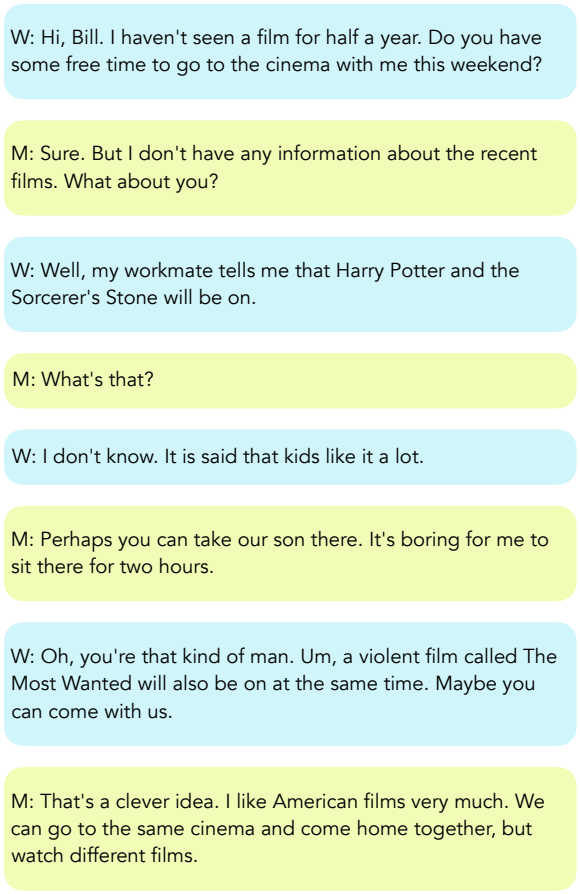


Figure 2: An example dialogue from the DREAM dataset (Sun et al., 2019).

summarization annotated in the DIALOGSUM dataset is the following:

"Person1 invites Bill to go to the cinema together this weekend. Person1 hears the Harry Potter movie would be on but Person2 likes the violent film."

From this comparison between the traditional free-form abstractive dialogue summarization and our proposed structured dialogue summarization, we can clearly see that the STRUDEL summarization includes more important aspects about the dialogue and tells a more comprehensive and informative story compared to the traditional free-form abstractive dialogue summarization.

4 Human Annotations of STRUDEL

Our proposed new task of Structured Dialogue Summarization (STRUDEL) opens up a gateway for language models to observe, imitate and learn from the structured human mental process of systematic dialogue understanding. But in order to actually infuse these valuable human-guided structural priors regarding dialogue understanding into language models through the task of STRUDEL, we first need to collect high-quality supervision information from empirical human demonstration of performing the STRUDEL task. Therefore, for this purpose, we collect a set of human annotations of STRUDEL over 400 dialogues sampled from two widely used dialogue comprehension datasets - the MuTual (Cui et al., 2020) dataset for dialogue response prediction and the DREAM (Sun et al., 2019) dataset for dialogue question answering. In our collection of STRUDEL human annotations, each sampled dialogue is manually annotated with its complete set of STRUDEL summarization with all 16 STRUDEL entries (can contain `'N/A'`) by a human annotator following the annotation protocols (see Section 4.2).

4.1 Datasets

The two dialogue comprehension datasets that we used for the human annotations of STRUDEL are:

4.1.1 MuTual

MuTual (Cui et al., 2020) is a popular recently proposed multi-turn dialogue reasoning dataset in the form of dialogue response prediction. All dialogue corpora in the MuTual dataset are modified from Chinese high school English listening comprehension test data, where students are expected to select the best answer from three candidate options, given a multi-turn dialogue and a question. Authors asked human annotators to rewrite the question and answer candidates as response candidates to fit in the test scenario of dialogue response prediction. MuTual consists of 8860 challenging questions. Almost all questions involve reasoning, which are designed by linguist experts and high-quality annotators. MuTual is the first human-labeled reasoning-based dataset for multi-turn dialogue.

4.1.2 DREAM

DREAM (Sun et al., 2019) is the first multiple-choice reading comprehension dataset on dialogues. It is collected from English comprehension examinations designed by human experts and contains 10197 multiple-choice questions for 6444 dialogues. DREAM presents a challenging in-depth, multi-turn and multi-party dialogue understanding task because of its features of being mostly non-extractive, requiring reasoning beyond single sen-

STRUDEL	DREAM		MuTual	
Entry Name	%	Avg Len	%	Avg Len
Name$_{S_1}$	31.5%	1.07	21%	1.06
Name$_{S_2}$	54.5%	1.18	45%	1.13
Role/ Identity$_{S_1}$	70%	1.22	57%	1.26
Role/ Identity$_{S_2}$	73.5%	1.21	58%	1.27
Relationship	72%	1.92	54.5%	1.69
Time	10.5%	1.09	6%	1.06
Location$_{S_1}$	36.5%	1.35	22.5%	1.22
Location$_{S_2}$	36 %	1.39	21.5%	1.21
Purpose/ Theme	100%	7.80	100 %	6.55
Task/ Intention$_{S_1}$	99.5%	7.36	99.5%	7.08
Task/ Intention$_{S_2}$	99%	7.10	97%	6.51
Problem/Dis agreement$_{S_1}$	94.5%	10.76	91.5%	9.97
Solution$_{S_1}$	92.5%	12.33	90.5%	11.25
Problem/Dis agreement$_{S_2}$	59%	6.22	46%	5.235
Solution$_{S_2}$	57 %	7.7	42%	5.38
Conclusion/ Agreement	90%	14.83	88%	15.74

Table 1: Statistics of our collected STRUDEL human annotations over the DREAM dataset and the MuTual dataset. '**%**' denotes frequency of appearance in percentage, and '**Avg Len**' denotes average length of each STRUDEL summarization entry as measured in number of words.

tences and involving commonsense knowledge.

4.2 Annotation Protocols

We use the JSON format for the manual annotation of STRUDEL. The two major annotation protocols we prescribed to the annotators during STRUDEL human annotation are:

1. When writing each STRUDEL summarization entry for a dialogue, please be informative, succinct, faithful and to the point.

2. When you think a certain STRUDEL entry can't be inferred from the dialogue or is not mentioned in the dialogue at all or doesn't apply to the current dialogue, please write `'N/A'` for that STRUDEL entry in your annotation.

See Figure 5 in Appendix A for an example human annotation of STRUDEL in JSON format.

4.3 Annotation Statistics

The statistics of our collected human annotations of STRUDEL are reported in Table 1.

5 Modeling Approach

In this section, we describe our main modeling approach that uses Structured Dialogue Summarization (STRUDEL) to improve pre-trained language model's ability of dialogue comprehension.

5.1 STRUDEL as a Meta-Model

As we can see from the definition in Section 3.1, Structured Dialogue Summarization (STRUDEL) is a generic task that can be generally applied to any dialogue. Therefore, STRUDEL can be viewed as an important upstream auxiliary NLU task and can be used to train language models to better understand dialogues in a structured and systematic way before they were further finetuned over specific downstream dialogue comprehension tasks.

As a result, based on our definition of STRUDEL, we further propose a new modeling framework of STRUDEL dialogue comprehension, in which STRUDEL can be viewed as a meta-model that can be smoothly integrated into and used on top of a wide range of different large-scale pre-trained transformer encoder models for dialogue understanding. Figure 1 provides a conceptual illustation of this relationship between STRUDEL and pre-trained language models. Below we discuss each of the different components of our STRUDEL dialogue comprehension framework in details.

5.2 STRUDEL Prompt Questions

We first design a prompt question for each STRUDEL summarization entry, which will be used to query a pre-trained language model to generate a vector embedding of that STRUDEL entry for a dialogue. For each STRUDEL summarization entry defined in Section 3.1, we add the common prefix `'Summarize: what is '` to its definition sentence and replace the `'.'` at the end with `'?'` to form its corresponding STRUDEL prompt question. For example, for STRUDEL entry (e), the *relationship* entry, its definition sentence is `'the relationship between the two speakers of the dialogue.'`, and its corresponding STRUDEL prompt question is `'Summarize: what is`

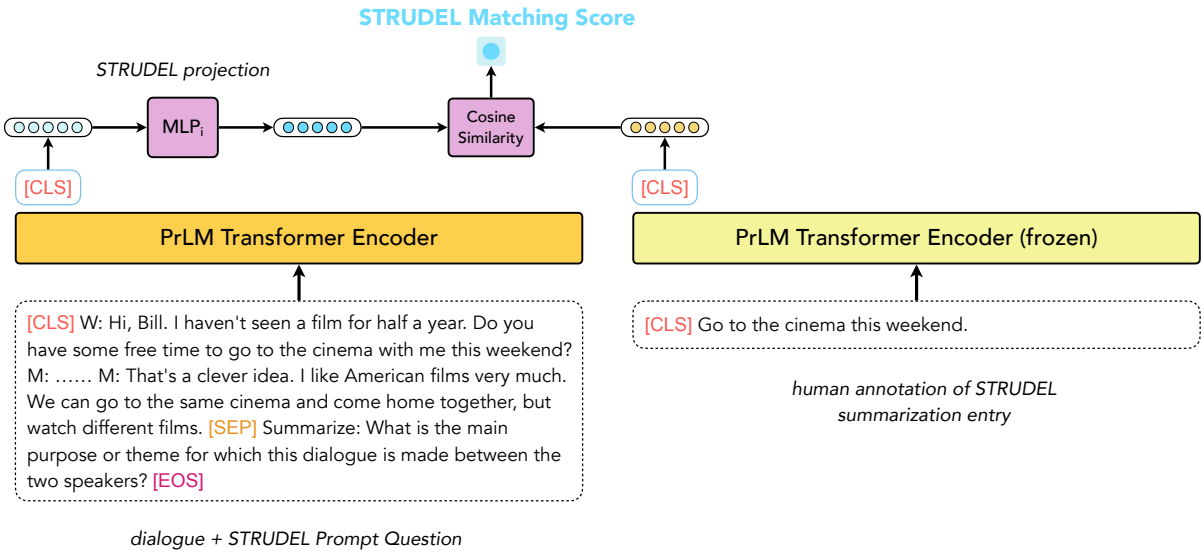


Figure 3: The modeling pipeline that trains a transformer encoder to learn to generate vector embeddings of STRUDEL entries that match their corresponding human annotations.

the relationship between the two speakers of the dialogue?'

5.3 Learning to Generate STRUDEL Embeddings

In our STRUDEL dialogue comprehension modeling framework, we choose to train transformer encoder language models to learn to generate semantic vector embeddings of the contents of STRUDEL entries instead of the actual text outputs of the STRUDEL entries in the form of token sequences. We make this design choice mainly for two reasons: (1) the form of vector embeddings makes it easier to quantitatively compare model-generated structured dialogue summarizations with their corresponding human annotations (e.g. by calculating cosine similarities in the vector space); (2) vector embeddings of STRUDEL can also be smoothly integrated back into transformer encoders for running inference over dialogue comprehension tasks.

Now we describe the procedure to train a pre-trained transformer encoder language model to learn to generate STRUDEL embeddings under the supervision from STRUDEL human annotations. Given a dialogue input sequence D and a pre-trained transformer encoder language model $\mathcal{T}$ for computing deep contextualized representations of textual sequences, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020) and ELECTRA (Clark et al., 2020), for an entry $\mathcal{E}$ of the STRUDEL summarization, we first concatenate D with the STRUDEL prompt question $Q_{\mathcal{E}}$ for the STRUDEL entry $\mathcal{E}$ (as defined in Section 5.2) together to form a query sequence $\{\texttt{[CLS]}\ D\ \texttt{[SEP]}\ Q_{\mathcal{E}}\ \texttt{[EOS]}\}$, and then feed this query sequence into the transformer encoder $\mathcal{T}$ to compute its contextualized representation. Let $H^{\mathcal{E}}$ be the last layer of hidden state vectors computed from this transformer encoder $\mathcal{T}$, then we have:

$$H^{\mathcal{E}} = \mathcal{T}\Big(\{\texttt{[CLS]}\ D\ \texttt{[SEP]}\ Q_{\mathcal{E}}\ \texttt{[EOS]}\}\Big) \tag{1}$$

Let $h^{\mathcal{E}}_{\texttt{[CLS]}}$ denote the last-layer hidden state vector of the [CLS] token in $H^{\mathcal{E}}$, then we apply a dedicated multi-layer perceptron $\text{MLP}^{\mathcal{E}}$ on top of $h^{\mathcal{E}}_{\texttt{[CLS]}}$ to project it onto a same-dimensional vector space to obtain our final vector embedding of the STRUDEL entry $\mathcal{E}$.

Now let $A^{\mathcal{E}}$ denotes the human-annotated ground-truth summarization for STRUDEL entry $\mathcal{E}$. Then we use a frozen version of the same transformer encoder, denoted as $\tilde{\mathcal{T}}$, to encode this human annotation as:

$$\tilde{H}^{\mathcal{E}} = \tilde{\mathcal{T}}\Big(\{\texttt{[CLS]}\ A^{\mathcal{E}}\}\Big) \tag{2}$$

Let $\tilde{h}^{\mathcal{E}}_{\texttt{[CLS]}}$ denote the last-layer hidden state vector of the [CLS] token in $\tilde{H}^{\mathcal{E}}$, then we can compute the semantic matching score between the transformer model's generated vector embedding for STRUDEL entry $\mathcal{E}$ and its corresponding human annotation as: $\text{Cos}\Big(\text{MLP}^{\mathcal{E}}(h^{\mathcal{E}}_{\texttt{[CLS]}}),\ \tilde{h}^{\mathcal{E}}_{\texttt{[CLS]}}\Big)$.

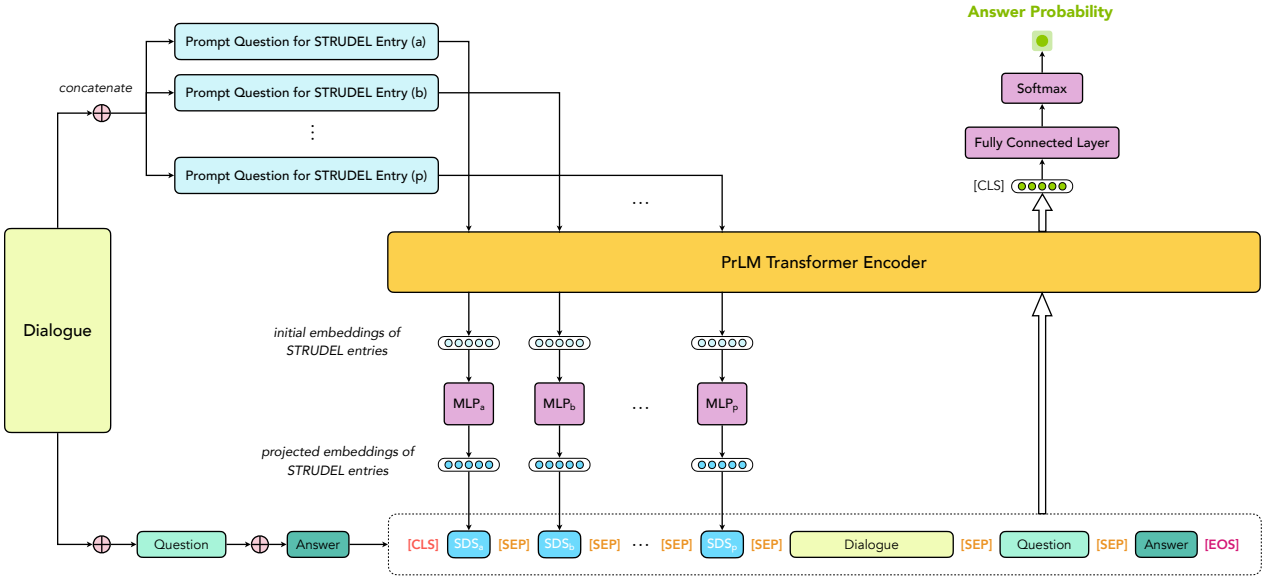


Figure 4: The overall model architecture of our STRUDEL dialogue comprehension modeling framework.

Therefore, the objective function for optimizing the transformer encoder model $\mathcal{T}$ to generate STRUDEL summarizations that matches human annotations can be formulated as:

$$\mathbb{L}_{\text{SM}} = -\sum_{\mathcal{E}\in\mathcal{S}} \text{Cos}\Big(\text{MLP}^{\mathcal{E}}(h^{\mathcal{E}}_{\texttt{[CLS]}}),\ \tilde{h}^{\mathcal{E}}_{\texttt{[CLS]}}\Big) \tag{3}$$

where $\mathcal{S}$ denotes the set of all 16 different STRUDEL entries. See Figure 3 for an illustration of this modeling pipeline.

5.4 STRUDEL for Dialogue Comprehension

After a transformer encoder language model learns to generate embeddings of structured dialogue summarization, we need to design a modeling framework to employ these generated STRUDEL embeddings to improve the model's dialogue comprehension ability. Here we focus on two important types of dialogue comprehension tasks - dialogue question answering and dialogue response prediction (Zhang and Zhao, 2021).

Given a dialogue input sequence D, a question Q, a candidate answer A (for dialogue response prediction tasks, Q will be empty and A will be a candidate response) and a transformer encoder language model $\mathcal{T}$, for each entry $\mathcal{E}$ of the STRUDEL summarization, we define a special STRUDEL token $\texttt{[SDS}_{\mathcal{E}}\texttt{]}$ to store the vector embedding of that STRUDEL entry $\mathcal{E}$ generated by the model $\mathcal{T}$. Then we append all the 16 STRUDEL tokens to the front of D to form an input sequence: I_{SDS} = $\{$ [CLS] $[\text{SDS}_a]$ [SEP] $[\text{SDS}_b]$ [SEP] $[\text{SDS}_c]$ ... [SEP] $[\text{SDS}_p]$ [SEP] D [SEP] Q [SEP] A [EOS] $\}$, and feed this sequence back to $\mathcal{T}$ to compute its last layer of contextualized representation as:

$$H^{\text{SDS}} = \mathcal{T}(I_{\text{SDS}}) \tag{4}$$

Let $h^{\text{SDS}}_{\texttt{[CLS]}}$ denote the last-layer hidden state vector of the [CLS] token in H, then we apply a fully connected layer followed by a softmax function on $h^{\text{SDS}}_{\texttt{[CLS]}}$ to compute the probability of the answer (or response) being the candidate A given the dialogue D and the question Q as:

$$\mathbb{P}^{\text{SDS}}(A \mid D, Q) = \text{Softmax}\Big(\text{FC}(h^{\text{SDS}}_{\texttt{[CLS]}})\Big) \tag{5}$$

Let a^* denote the correct answer (or response) in the training labels, then the objective function that we use to train the transformer encoder language model $\mathcal{T}$ to use STRUDEL summarization embeddings to perform dialogue question answering (or response prediction) can be formulated as the cross-entropy loss:

$$\mathbb{L}_{\text{CE}} = -\log\Big(\mathbb{P}^{\text{SDS}}(A = a^* \mid D, Q)\Big) \tag{6}$$

See Figure 4 for an illustration of the above model architecture for STRUDEL dialogue comprehension.

5.5 Model Training

5.5.1 Multi-Task Post-Training

During the training of our STRUDEL dialogue comprehension model, we first adopt a multi-task

Model	MuTual			DREAM
	$\mathbf{R}_4$@1	$\mathbf{R}_4$@2	MRR	Accuracy
RoBERTa$_{large}$ (Liu et al., 2019)	0.695	0.878	0.824	0.821
RoBERTa$_{large}$ + STRUDEL	**0.869**	**0.947**	**0.919**	**0.838**
ALBERT$_{large}$ (Lan et al., 2020)	0.656	0.853	0.796	0.568
ALBERT$_{large}$ + STRUDEL	**0.673**	**0.872**	**0.812**	**0.596**

Table 2: Our experiment results on the MuTual dataset and the DREAM dataset.

learning strategy to train the transformer model to learn to generate accurate STRUDEL embeddings and to infer the correct choices for dialogue question answering and response prediction tasks based on its generated STRUDEL embeddings at the same time. This multi-task post-training process uses distinct but complementary tasks to challenge the model to learn structured and meaningful representations of dialogue semantics that is widely generalizable to different dialogue comprehension tasks. To do this, we define our objective function to be an average of the weighted sum of the semantic matching loss defined in Equation 3 and the cross-entropy loss defined in Equation 6:

$$\mathbb{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\alpha_1 \mathbb{L}^i_{\text{SM}} + \alpha_2 \mathbb{L}^i_{\text{CE}}\right) \tag{7}$$

where N is the total number of dialogue examples.

5.5.2 Single-Task Fine-Tuning

After our transformer-based STRUDEL dialogue comprehension model has been post-trained using the objective function defined in Equation 7, we take the model checkpoint and continue to fine-tune the model over individual dialogue comprehension tasks in order to fully maximize its performance on each of the tasks.

6 Experiments

6.1 Transformer Encoder Models

In our experiment, we use two widely-used transformer encoder language models - RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020) - as the backbone transformer encoder in our STRUDEL dialogue comprehension modeling framework.

6.2 Dialogue Comprehension Tasks

In our experiment, we test our STRUDEL dialogue comprehension model on two important and representative dialogue comprehension tasks - dialogue question answering and dialogue response prediction. We use the DREAM dataset and the MuTual dataset introduced in Section 4.1 to train and test our model over the two tasks respectively.

6.3 Results

The results of our experiments are shown in Table 2. As we can see from the table, the accuracy results of our STRUDEL dialogue comprehension models on both the dialogue response prediction task (over the MuTual dataset) and the dialogue question answering task (over the DREAM dataset) are all consistently higher than their corresponding backbone transformer encoder models alone. This clearly demonstrates that our proposed task of Structured Dialogue Summarization (STRUDEL) and our proposed STRUDEL dialogue comprehension modeling framework can indeed help transformer language models to learn to better perform dialogue comprehension tasks.

7 Conclusion

In this paper, we presented STRUDEL (STRUctured DiaLoguE Summarization) - a novel type of dialogue summarization task that can help pre-trained language models to better understand dialogues and improve their performance on important dialogue comprehension tasks. In contrast to the traditional free-form abstractive summarization task for dialogues, STRUDEL provides a more comprehensive digest over multiple important aspects of a dialogue and has the advantage of being more focused, specific and instructive for dialogue comprehension models to learn from. In addition, we also introduced a new STRUDEL dialogue comprehension modeling framework that

integrates STRUDEL into a dialogue reasoning module over transformer encoder language models to improve their dialogue comprehension ability. Our empirical experiments on the tasks of dialogue question answering and dialogue response prediction confirmed that our STRUDEL dialogue comprehension modeling framework can significantly improve the dialogue comprehension performance of transformer encoder language models.

8 Limitations

There are two major limitations of our work discussed in this paper:

1. Our paper mainly focuses on designing the structured dialogue summarization task for two-speaker dialogues, which is the majority of multi-turn dialogues that are most commonly seen in dialogue datasets and real applications. In the future, we plan to further extend our STRUDEL framework to also accommodate multi-speaker dialogues between more than two speakers.

2. Our approach haven't included any explicit knowledge reasoning components yet, which are also important for language models to accurately generate structured dialogue summarizations and perform dialogue comprehension tasks. In future work, we plan to integrate a knowledge reasoning module into our STRUDEL dialogue summarization modeling framework in order to further improve its performance.

Acknowledgment

We would like to thank Meta AI for their generous support of our work.

References

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

JianCheng Du and Yang Gao. 2021. Query-focused abstractive summarization via question-answering model. In *2021 IEEE International Conference on Big Knowledge (ICBK)*, pages 440–447.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *CoRR*, abs/2107.03175.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2020a. Dialogue discourse-aware graph model and data augmentation for meeting summarization.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2020b. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. *CoRR*, abs/2010.10044.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*.

Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2020. Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. *CoRR*, abs/2009.06504.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. *CoRR*, abs/2106.01890.

P Mahalakshmi and N Sabiyath Fatima. 2022. Summarization of text and image captioning in information retrieval using deep learning techniques. *IEEE Access*, 10:18289–18297.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2020. Dialogue graph modeling for conversational machine reading. *CoRR*, abs/2012.14827.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation.

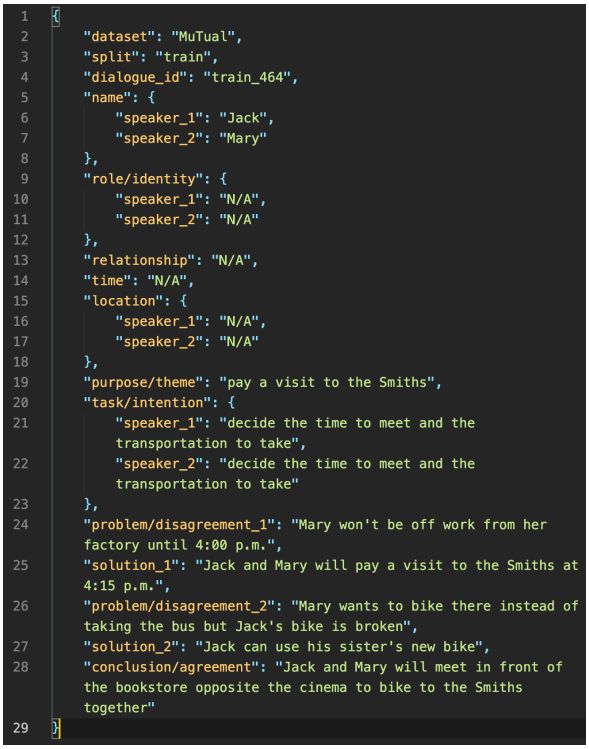
Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Johannes Villmow, Adrian Ulges, and Ulrich Schwanecke. 2021. A structural transformer with relative positions in trees for code-to-sequence tasks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE.

Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Ziqiang Cao, Sujian Li, Hua Wu, and Haifeng Wang. 2021. Bass: Boosting abstractive summarization with unified semantic graph. *arXiv preprint arXiv:2105.12041*.

Zhuosheng Zhang and Hai Zhao. 2021. Advances in multi-turn dialogue comprehension: A survey.

A Example of STRUDEL Human Annotations



```
{
    "dataset": "MuTual",
    "split": "train",
    "dialogue_id": "train_464",
    "name": {
        "speaker_1": "Jack",
        "speaker_2": "Mary"
    },
    "role/identity": {
        "speaker_1": "N/A",
        "speaker_2": "N/A"
    },
    "relationship": "N/A",
    "time": "N/A",
    "location": {
        "speaker_1": "N/A",
        "speaker_2": "N/A"
    },
    "purpose/theme": "pay a visit to the Smiths",
    "task/intention": {
        "speaker_1": "decide the time to meet and the transportation to take",
        "speaker_2": "decide the time to meet and the transportation to take"
    },
    "problem/disagreement_1": "Mary won't be off work from her factory until 4:00 p.m.",
    "solution_1": "Jack and Mary will pay a visit to the Smiths at 4:15 p.m.",
    "problem/disagreement_2": "Mary wants to bike there instead of taking the bus but Jack's bike is broken",
    "solution_2": "Jack can use his sister's new bike",
    "conclusion/agreement": "Jack and Mary will meet in front of the bookstore opposite the cinema to bike to the Smiths together"
}
```


Figure 5: An example human annotation of STRUDEL in JSON format for the dialogue 'train_464' from the MuTual dataset.