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. We further collect human annotations of STRUDEL summaries over 400 dialogues and introduce a new STRUDEL dialogue comprehension modeling framework that integrates STRUDEL into a graph-neural-network-based dialogue reasoning module over transformer encoder language models to improve their dialogue comprehension abilities. In our empirical experiments on two important downstream dialogue comprehension tasks - dialogue question answering and dialogue response prediction - we show that our STRUDEL dialogue comprehension model can significantly improve the dialogue comprehension performance of transformer encoder language models.

1 Introduction

In natural language processing, abstractive dialogue summarization (Chen et al., 2021; Gliwa et al., 2019; Fabbri et al., 2021; Zhu et al., 2021; Zhong 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
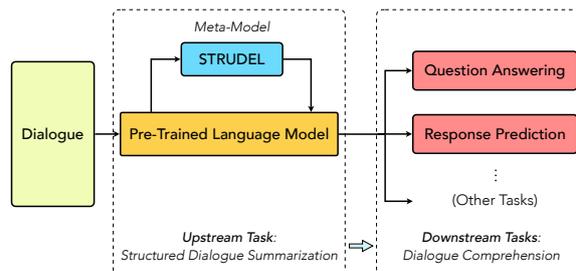

Figure 1: STRUDEL as a meta-model on top of pre-trained language models for dialogue comprehension.

the main content of a piece of text, usually from multiple perspectives each focusing on 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 involves a 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 better perform dialogue comprehension tasks?

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 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. And guided by our proposed

---

[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.

STRUDEL task, we also collect a set of human annotations of STRUDEL summaries over 400 different dialogues sampled from two widely used dialogue comprehension datasets, MuTual (Cui et al., 2020) and DREAM (Sun et al., 2019). Then we further introduce a new dialogue comprehension model that integrates STRUDEL into a dialogue reasoning module based on graph neural networks over transformer encoder language models. Our empirical experiment results show 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 Related Work

2.1 Abstractive Text Summarization

Abstractive text summarization (See et al., 2017; Rush et al., 2015; Zhang et al., 2020) is the task of generating a concise summary of a source text by paraphrasing its main content instead of simply extracting its important sentences, which is referred to as extractive summarization. Recently, new methods of abstractive summarization started to take the structure of the source document into consideration when generating a single free-form abstractive summarization. For example, Balachandran et al. (2021) proposed a new summarization framework based on document-level structure induction that incorporates latent and explicit dependencies across sentences in the source document into summarization models. And Wu et al. (2021) presented a novel framework for boosting abstractive summarization called BASS that is based on a unified semantic graph and a graph-based encoder-decoder model.

2.2 Abstractive Dialogue Summarization

Abstractive dialogue summarization (Chen et al., 2021; Gliwa et al., 2019; Fabbri et al., 2021; Zhu et al., 2021; Zhong et al., 2021) is a particular type of abstractive text summarization task that focuses on the abstractive summarization of human dialogues and conversations. Recently, there has been a lot of work that tries to improve the quality and faithfulness of model-generated abstractive dialogue summaries. For example, Feng et al. (2020) proposed to use heterogeneous graph networks to incorporate commonsense knowledge into

abstractive dialogue summarization models. And Tang et al. (2022) proposed to improve the quality and faithfulness of abstractive dialogue summarization through a new training strategy based on linguistically-informed contrastive fine-tuning.

2.3 Dialogue Comprehension and Understanding

There have also been many advances in multi-turn dialogue comprehension and understanding (Zhang and Zhao, 2021a) in recent years. For example, Liu et al. (2020) showed that explicitly modeling utterance-aware and speaker-aware representations can boost the dialogue comprehension performance of pre-trained language models. Ouyang et al. (2020) showed that segmenting a dialogue into elementary discourse units (EDUs) and then use a graph modeling framework to model their relationships can help language models to better understand the dialogue's innate structure. Zhang and Zhao (2021b) proposed a new structural pre-training method for dialogue comprehension that captures dialogue-exclusive features by adding utterance order restoration and sentence backbone regularization into the language modeling objectives. And neural-retrieval-in-the-loop architectures have been shown to reduce hallucination in conversation models (Shuster et al., 2021).

3 Structured Dialogue Summarization

In this section we introduce our proposed new task of structured dialogue summarization, give its detailed definition and demonstrate its features through a concrete example.

3.1 Definition of Structured Dialogue Summarization

We define *Structured Dialogue Summarization* (STRUDEL) as the task of generating a systematic and abstractive 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 7 different STRUDEL entries, which are each defined as follows:

---

[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).

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

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

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

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

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

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

(g) **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 7 STRUDEL entries will either be a short text abstractively summarizing the 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:

Relationship: *"Wife and husband."*
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: *"It's boring for Bill to watch the film Happy Potter and the Sorcerer's Stone."*
Solution: *"Bill will watch another film called the Most Wanted."*

W: Hi, Bill. I haven't seen a film for half a year. Do you have some free time to go to the cinema with me this weekend?
M: Sure. But I don't have any information about the recent films. What about you?
W: Well, my workmate tells me that Harry Potter and the Sorcerer's Stone will be on.
M: What's that?
W: I don't know. It is said that kids like it a lot.
M: Perhaps you can take our son there. It's boring for me to sit there for two hours.
W: Oh, you're that kind of man. Um, a violent film called The Most Wanted will also be on at the same time. Maybe you can come with us.
M: That's a clever idea. I like American films very much. We can go to the same cinema and come home together, but watch different films.

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

Conclusion/Agreement: *"Go to the cinema and come home together, but watch different films."*

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

STRUDEL	DREAM		MuTual	
Entry Name	%	Avg Len	%	Avg Len
Relationship	72%	1.92	54.5%	1.69
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/Disagreement	94.5%	10.76	91.5%	9.97
Solution	92.5%	12.33	90.5%	11.25
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 (i.e. not annotated with 'N/A') in percentage, and '**Avg Len**' denotes average length of each STRUDEL summarization entry as measured by number of words.

we first need to collect high-quality supervision information from the 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 dataset (Cui et al., 2020) for dialogue response prediction and the DREAM dataset (Sun et al., 2019) 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 7 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. The 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 has a total of 8860 expert-designed challenge questions, almost all of which require reasoning.

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 different 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 sentences, and involving commonsense knowledge.

4.2 Annotation Protocols

The two major annotation protocols that we prescribed to the annotators during the STRUDEL human annotation process are:

1. When writing each STRUDEL summarization entry for a dialogue, please be informative, concise, 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.

4.3 Annotation Statistics

The statistics of our collected human annotations of STRUDEL are reported in Table 1. From these statistics, we can see that the majority of STRUDEL entries have high appearance frequency across dialogues, and on average the annotated STRUDEL entries are relatively concise.

5 Modeling Approach

In this section, we describe our main modeling approach that uses Structured Dialogue Summarization (STRUDEL) to improve pre-trained language models' abilities 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
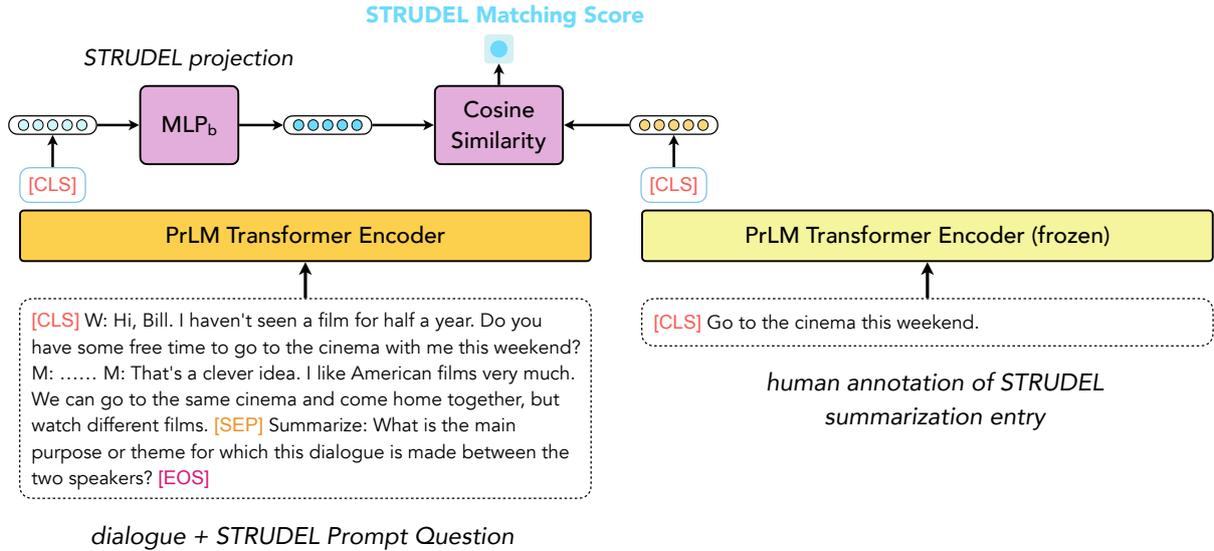
Figure 3: The modeling pipeline that trains a transformer encoder to learn to generate vector embeddings of STRUDEL entries that match their corresponding human annotations.

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 fine-tuned 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 illustration of this relationship between STRUDEL and pre-trained language models. Below we discuss each of the different components of our STRUDEL dialogue comprehension modeling framework in detail.

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 (a), 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 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 textual 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 into a reasoning module based on graph neural networks 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 of 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) and RoBERTa (Liu et al., 2019), 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:

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

Let $h^{\mathcal{E}}_{\texttt{[CLS]}}$ denote the last-layer hidden state vector of the $\texttt{[CLS]}$ 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 $A^{\mathcal{E}}$ denote 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:

$$\tilde{H}^{\mathcal{E}} = \tilde{\mathcal{T}}\Big(\big\{\texttt{[CLS]}\,A^{\mathcal{E}}\,\texttt{[EOS]}\big\}\Big)$$

Let $\tilde{h}^{\mathcal{E}}_{\texttt{[CLS]}}$ denote the last-layer hidden state vector of the $\texttt{[CLS]}$ 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 their cosine similarity value: $\cos\Big(\text{MLP}^{\mathcal{E}}(h^{\mathcal{E}}_{\texttt{[CLS]}}),\ \tilde{h}^{\mathcal{E}}_{\texttt{[CLS]}}\Big)$. Therefore, the objective function for optimizing the transformer encoder model $\mathcal{T}$ to generate STRUDEL summaries that match human annotations can be formulated as:

$$\mathbb{L}_{\text{SM}} = -\sum_{\mathcal{E}} \cos\Big(\text{MLP}^{\mathcal{E}}(h^{\mathcal{E}}_{\texttt{[CLS]}}),\ \tilde{h}^{\mathcal{E}}_{\texttt{[CLS]}}\Big) \tag{1}$$

See Figure 3 for an illustration of this modeling pipeline.

5.4 Dialogue Semantic Graph

In order to systematically organize all the information contained in different STRUDEL summary entries into a structured representation that is amenable to collective reasoning and inference,
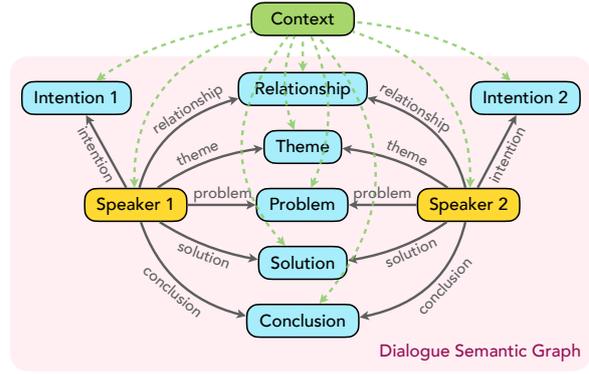

Figure 4: The dialogue semantic graph of STRUDEL embeddings and the context node on top of the graph. The speaker nodes are depicted in color yellow and the STRUDEL embedding nodes are depicted in color blue.

we propose to construct a graph representation of STRUDEL embeddings called **dialogue semantic graph**. The dialogue semantic graph consists of two types of nodes: 2 speaker nodes (one for the first speaker and the other one for the second speaker) and 7 STRUDEL embedding nodes (one for each of the 7 STRUDEL summary entries). It also has 6 different types of relational edges, each corresponding to one of the following 6 different relations from speakers to STRUDEL summaries: '*has_relationship*', '*shares_theme*', '*with_intention*', '*faces_problem*', '*finds_solution*', and '*reaches_conclusion*'. The structure of the dialogue semantic graph is depicted in Figure 4.

5.5 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 abilities. In this paper we focus on two important types of dialogue comprehension tasks - dialogue question answering and dialogue response prediction (Zhang and Zhao, 2021a). Since all the generated STRUDEL embeddings can be organized into a dialogue semantic graph, we can build a graph neural network (GNN) module on top of the dialogue semantic graph to perform reasoning for different dialogue comprehension tasks. In this paper we adopt the QA-GNN (Yasunaga et al., 2021) architecture as our GNN reasoning module to perform context-aware structured reasoning over generated STRUDEL embeddings.
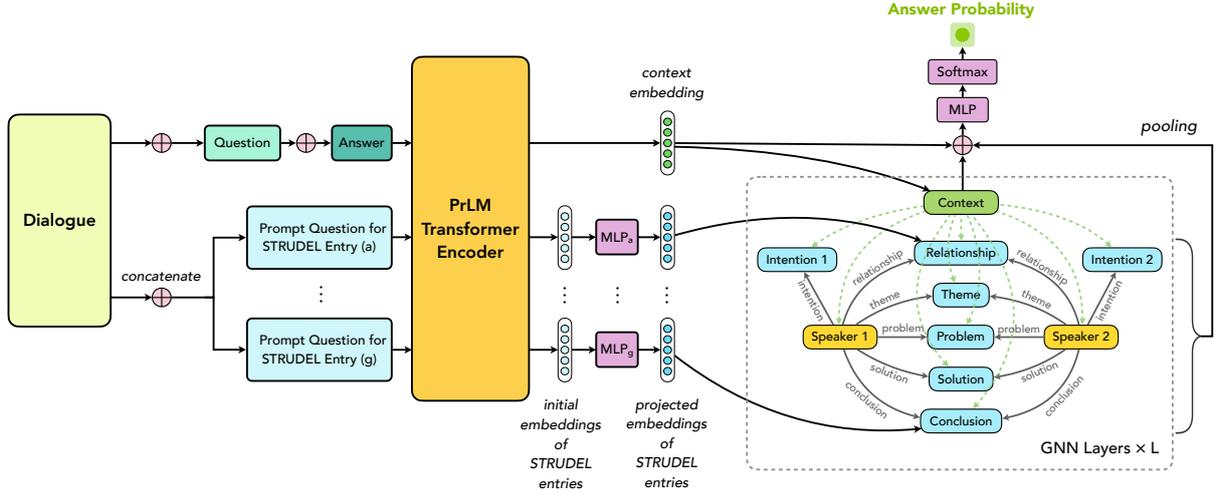
Figure 5: The overall model architecture of our STRUDEL dialogue comprehension modeling framework.

More specifically, 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 $\mathcal{T}$, we first use the procedure described in Section 5.3 to compute the STRUDEL embedding vector $v^{\mathcal{E}}$ for each STRUDEL entry $\mathcal{E}$ as: $v^{\mathcal{E}} = \mathrm{MLP}^{\mathcal{E}}(h^{\mathcal{E}}_{[\mathrm{CLS}]})$. Then we form a context sequence $\{[\mathrm{CLS}]\,D\,[\mathrm{SEP}]\,Q\,[\mathrm{SEP}]\,A\,[\mathrm{EOS}]\}$ and feed it into the same transformer encoder $\mathcal{T}$ to compute its contextualized representation:

$$H = \mathcal{T}\Big(\{[\mathrm{CLS}]\,D\,[\mathrm{SEP}]\,Q\,[\mathrm{SEP}]\,A\,[\mathrm{EOS}]\}\Big).$$

Let $h_{[\mathrm{CLS}]}$ denote the last-layer hidden state vector of the $[\mathrm{CLS}]$ token in H, then we take $h_{[\mathrm{CLS}]}$ as the context embedding vector $v^{\mathcal{C}}$ that encodes the original dialogue, the question and the current answer candidate (for dialogue response prediction tasks it encodes the original dialogue and the current response option): $v^{\mathcal{C}} = h_{[\mathrm{CLS}]}$. And additionally, for the two speakers $\mathcal{S}_1$ and $\mathcal{S}_2$ of the dialogue, we use the frozen transformer encoder $\tilde{\mathcal{T}}$ to compute the following two contextualized representations:

$$H^{\mathcal{S}_1} = \tilde{\mathcal{T}}\Big(\{\text{`\textit{The first speaker of this dialogue.}'}\}\Big),$$

$$H^{\mathcal{S}_2} = \tilde{\mathcal{T}}\Big(\{\text{`\textit{The second speaker of this dialogue.}'}\}\Big),$$

and assign the embedding vectors of the two speakers to be the last-layer hidden state vectors of the corresponding $[\mathrm{CLS}]$ tokens: $v^{\mathcal{S}_1} = h^{\mathcal{S}_1}_{[\mathrm{CLS}]}$ and $v^{\mathcal{S}_2} = h^{\mathcal{S}_2}_{[\mathrm{CLS}]}$.

Following the architecture of QA-GNN (Yasunaga et al., 2021), to build our GNN reasoning module over STRUDEL embeddings, we add a context node on top of the dialogue semantic graph and add 9 edges between the context node and all the 9 nodes in the dialogue semantic graph to form a joint graph for dialogue reasoning (see Figure 4). Then we construct an L-layer graph attention network (GAT) (Veličković et al., 2018) using this joint graph and integrate it with the central transformer encoder language model $\mathcal{T}$ to perform reasoning over dialogue comprehension tasks. More specifically, we first add dimension projection layers to project the context embedding vector $v^{\mathcal{C}}$, the two speaker embedding vectors $v^{\mathcal{S}_1}$ and $v^{\mathcal{S}_2}$, and all the STRUDEL embedding vectors $v^{\mathcal{E}}$ to the dimensionality of the QA-GNN node embeddings. Then we use the dimension-projected context embedding vector $\hat{v}^{\mathcal{C}}$ to initialize the context node, use the two dimension-projected speaker embedding vectors $\hat{v}^{\mathcal{S}_1}$ and $\hat{v}^{\mathcal{S}_2}$ to initialize the two speaker nodes, and use all the dimension-projected STRUDEL embedding vectors $\hat{v}^{\mathcal{E}}$ to initialize the corresponding STRUDEL entry nodes in our joint graph.

And similar to QA-GNN, after L-layers of GNN message passing over the joint graph, we concatenate the context embedding vector, the last-layer representation vector of the context node and the pooling of the last-layer representation vectors of all the nodes in the dialogue semantic graph together into a single vector and feed it into another multi-layer perceptron to obtain the final probability score for (D, Q, A). We then apply the softmax function over the probability scores of all the dif-

Model	MuTual			DREAM
	$\mathbf{R}_4@1$	$\mathbf{R}_4@2$	MRR	Accuracy
BERT-large (Devlin et al., 2019)	0.662	0.858	0.803	0.636
BERT-large + STRUDEL	**0.727**	**0.881**	**0.839**	**0.668**
RoBERTa-large (Liu et al., 2019)	0.786	0.931	0.879	0.821
RoBERTa-large + STRUDEL	**0.887**	**0.969**	**0.937**	**0.839**

Table 2: Our experiment results on the MuTual dataset (Cui et al., 2020) and the DREAM dataset (Sun et al., 2019). Following Cui et al. (2020), we report recall at position 1 in 4 candidates ($\mathbf{R}_4@1$), recall at position 2 in 4 candidates ($\mathbf{R}_4@2$) and Mean Reciprocal Rank (MRR) for the models' performance on the MuTual dataset. The higher value in each comparison pair is highlighted in **bold**.

ferent answer (or response) options to obtain their inferred probability distribution as:

$$\mathbb{P}(A \mid D, Q) = \text{Softmax}\Big(\text{QA-GNN}(v^{\mathcal{C}}, v^{\mathcal{S}_1}, v^{\mathcal{S}_1}, \{v^{\mathcal{E}}\})\Big)$$

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

$$\mathbb{L}_{\text{CE}} = -\log\Big(\mathbb{P}(A = a^* \mid D, Q)\Big) \quad (2)$$

See Figure 5 for an illustration of the overall model architecture for STRUDEL dialogue comprehension described above.

5.6 Model Training

5.6.1 Multi-Task Post-Training

During the training of our STRUDEL dialogue comprehension model, we first adopt a multi-task learning strategy to train our 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 are widely generalizable to different dialogue comprehension tasks. To do this, we define our objective function to be a weighted sum of the average of the semantic matching loss (defined in Equation 1) across all the human-annotated dialogues and the average of the cross-entropy loss (defined in Equation 2) across all the dialogue comprehension examples:

$$\mathbb{L} = \frac{1}{N_{ha}} \sum_{i=1}^{N_{ha}} \Big(\alpha_1 \mathbb{L}_{\text{SM}}^i\Big) + \frac{1}{N} \sum_{j=1}^{N} \Big(\alpha_2 \mathbb{L}_{\text{CE}}^j\Big) \quad (3)$$

where N is the total number of dialogue comprehension examples, N_{ha} is the total number of dialogues annotated with STRUDEL summaries in the human annotation process, and α_1 and α_2 are the hyperparameters that correspond to the weights in the weighted sum.

5.6.2 Single-Task Fine-Tuning

After our transformer-based STRUDEL dialogue comprehension model has been post-trained using the objective function defined in Equation 3, we take the post-trained 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 - BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) - 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 MuTual dataset (Cui et al., 2020) and the DREAM dataset (Sun et al., 2019) introduced in Section 4.1 to train and test our model.

6.3 Results

The results of our experiments are shown in Table 2. As we can see from the table, the performance scores 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 demonstrates that our proposed task of Structured Dialogue Summarization (STRUDEL) and our proposed STRUDEL dialogue comprehension modeling framework can indeed help transformer language models 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 GNN-based dialogue reasoning module over transformer encoder language models to improve their dialogue comprehension abilities. Our empirical experiments on the tasks of dialogue question answering and dialogue response prediction showed that our STRUDEL dialogue comprehension modeling framework can significantly improve the dialogue comprehension performance of transformer encoder language models.

8 Limitations

There are two 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 ap-

plications. In the future, we plan to further extend our STRUDEL framework to also accommodate multi-speaker dialogues among more than two speakers.

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

Acknowledgment

References

Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime Carbonell, and Yulia Tsvetkov. 2021. StructSum: Summarization via structured representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2575–2585, Online. Association for Computational Linguistics.

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.

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.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. 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), pages 6866–6880, Online. Association for Computational Linguistics.

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

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, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

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.

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

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

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.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In Proceedings of the 2022 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In International Conference on Learning Representations.

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.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 535–546, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 11328–11339. PMLR.

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

Zhuosheng Zhang and Hai Zhao. 2021b. Structural pre-training for dialogue comprehension. 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), pages 5134–5145, Online. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5905–5921, Online. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5927–5934, Online. Association for Computational Linguistics.