Transformer-based Multi-Party Conversation Generation using Dialogue Discourse Acts Planning

Alexander Chernyavskiy and **Dmitry Ilvovsky**
National Research University Higher School of Economics
Moscow, Russia
alschernyavskiy@gmail.com; dilvovsky@hse.ru

Abstract

Recent transformer-based approaches to multi-party conversation generation may produce syntactically coherent but discursively inconsistent dialogues in some cases. To address this issue, we propose an approach to integrate a dialogue act planning stage into the end-to-end transformer-based generation pipeline. This approach consists of a transformer fine-tuning procedure based on linearized dialogue representations that include special discourse tokens. The obtained results demonstrate that incorporating discourse tokens into training sequences is sufficient to significantly improve dialogue consistency and overall generation quality. The suggested approach performs well, including for automatically annotated data. Apart from that, it is observed that increasing the weight of the discourse planning task in the loss function accelerates learning convergence.

1 Introduction

The popularity of dialogue systems has resulted in an increased demand for their utilization in various applications. Existing approaches are largely focused on two-party conversations, which is applicable in chat-bots and assistance systems (Shang et al., 2015; Wang et al., 2015; Young et al., 2018; Gu et al., 2019). At the same time, there is another type of dialogue, known as multi-party conversations (Traum, 2003; Uthus and Aha, 2013; Ouchi and Tsuboi, 2016; Le et al., 2019). In this case, several interlocutors are involved in the dialogue, and the dialogue tree, consisting of successive utterances, is wide enough. This type of dialogue can be observed in Internet forum discussion threads.

Due to the complexity of the structure of MPC dialogues, it becomes more challenging for the base seq2seq models to generate response texts. Multi-task learning and external knowledge can be considered to simplify the utterance generation task. To this end, we additionally leverage the theory of dialogue acts (Stone et al., 2013; Zhang et al., 2017), which shows by which discourse rhetorical relations (more precisely, dialogue acts) the individual utterances of the dialogue are connected. Figure 1 shows an example of such a structure.

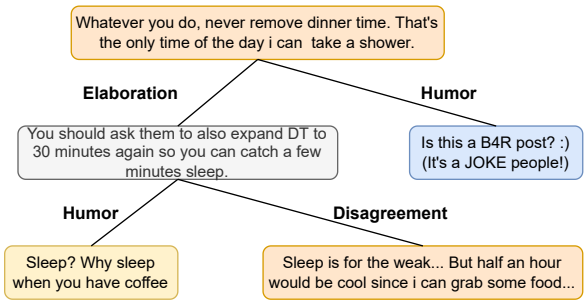

Figure 1: A manually annotated discourse tree for the multi-party dialogue. The color identifies the speaker.

Our major idea is that the use of dedicated discourse tokens to both input and target texts will enhance the coherence of discourse and consequently the overall quality of generation.

To illustrate, let us examine the following example of a help request forum thread. Initially, the user describes a problem and seeks advice. Subsequently, multiple dialogue turns occur, culminating in the following phrases:

- *[answer]* reinstall OS/get a new HDD/SSD
- *[disagreement]* really? My HDD was working right until yesterday...

Here, discourse relations demonstrate that the last utterance is indicative of a disagreement rather than a question. Accordingly, the next appropriate utterance should contain an inquiry to resolve the user's initial problem. Our discourse-based model generated "is your HDD in safe mode?", while the base model outputted "The answer is no.", which is much more distant from the corresponding ground-truth utterance, "how long have you had your PC?". This shows the advantage of the discourse-based

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 519–529
September 11–15, 2023.

model, as it first plans out discourse relations before generating text tokens.

Generally speaking, we suggest multi-task learning consisting of dialogue acts planning and response generation joined in the single pipeline. We integrate discourse tokens into a two-stage pipeline for MPC generation (see Section 3.1 for details). Its first stage is used to identify a speaker and an addressee at the current step, whereas the second stage is used to generate the current response text. The first part is quite challenging, but recent studies allow one to solve it qualitatively (Le et al., 2019; Gu et al., 2021). At the same time, only base models were researched for the second stage, leaving the relevance of discourse usage in dialogue generation unexplored. Therefore, we mainly focus on the second stage.

The task can be formalized as a graph2text, within which the BART and T5 models have already been partially investigated. Key part here is a linearization technique that was used for some graph structures (Ribeiro et al., 2020; Kale and Rastogi, 2020), but not for the discourse structure and dialogue generation yet. Thus, we suggest integrating dialogue acts into the linearization of MPC graphs.

Our contributions can be summarized as follows:

- We suggest multi-task learning consisting of dialogue acts planning and response generation to improve the transformer-based MPC generation pipeline.
- We analyze the importance of having discourse tokens in both parts of seq2seq linearized input pairs.
- We show that the transformer-based approach converges faster if it has more weight in the loss related to the dialogue acts planning task.

The code is available at `https://github.com/alchernyavskiy/discourse_mpc_generation`.

2 Related Work

In this paper, we consider multi-party conversations (MPCs). The process of generation is generally split into two stages since it consists of several entire tasks. Ouchi and Tsuboi (2016) presented a task of identifying the speaker and the addressee of an utterance (first stage), and recent approaches have been aimed at improving results in this task (Le et al., 2019; Gu et al., 2021).

At the second stage, associated with the response generation task, some approaches use GCN to encode the complex MPC structure (Hu et al., 2019). It intended to improve prior recurrent neural network-based sequence-to-sequence (seq2seq) generation models (Luan et al., 2016; Serban et al., 2017). At the same time, it was shown that recent transformer-based approaches, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), achieve top results in various generation tasks, including dialogue generation. Therefore, BART and T5 are commonly used as the base generation models in recent approaches. For instance, Li et al. (2021b) uses the BART as a backbone to train the model that considers long-range contextual emotional relationships.

Moreover, recent transformer-based approaches effectively solve graph-to-text generation tasks. Ribeiro et al. (2020) demonstrated that BART and T5 outperform various GNNs trained to encode AMR graphs in the AMR-to-text task. Similarly, Kale and Rastogi (2020) indicated that pre-training in the form of T5 enables simple, end-to-end models to outperform pipelined neural architectures tailored for data-to-text generation. The key factor here is that any graph can be linearized, and Hoyle et al. (2021) showed that transformers are invariant to the method by which graphs are linearized. Thus, we do not explore ways of linearizing dialogue graphs augmented by discourse relations, but choose one of the most reasonable ones.

Discourse parsing of multi-party conversations is an adjacent direction that is gaining popularity. There are several works where the dialogues were analyzed in terms of discourse structure ad discourse relations parsing (Afantenos et al., 2015; Shi and Huang, 2019; Wang et al., 2021; Koto et al., 2021). Despite this, there are not a large number of publicly available datasets with discursively-annotated dialogues. Basically, all comparisons are conducted for the STAC dataset (Asher et al., 2016), which is quite small. As far as we know, the only large publicly available dataset is the CDSC dataset (Zhang et al., 2017). The main difference in our research from this direction is that we do not aim to suggest a novel discourse parser. At the same time, we use existing parsers and explore the importance of using discourse in the applied generative task.

Our idea of generating discourse relations in dialogues comes from the story-telling task. To

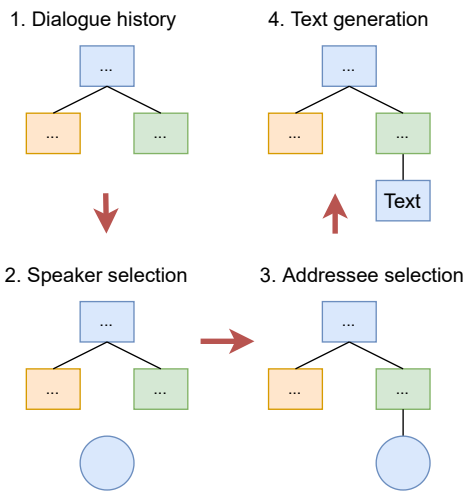

Figure 2: End-to-end MPC generation pipeline. The colors represent the speakers and are chosen as an example.

facilitate discourse coherence, some researchers proposed neural text generation based on discourse planning with an auxiliary model (Ji et al., 2016; Harrison et al., 2019; Chernyavskiy, 2022). However, in the case of dialogues, discourse structure has been explored only in the context of summarization and machine reading comprehension tasks (Feng et al., 2021; Li et al., 2021a).

3 Methods

3.1 End-to-End MPC Generation Pipeline

Figure 2 illustrates the end-to-end pipeline of multi-party conversation generation, which consists of several main steps at each dialogue turn. This pipeline implies that the next turn speaker selection can be separated from the response selection. There are also united approaches, but we do not consider them in this paper.

Firstly, the next speaker should be selected. Then, we should decide to which utterance it responds, or in other words, select the addressee of the generating utterance. Both these steps ($1 \rightarrow 2$ and $2 \rightarrow 3$ in Figure) are typically combined into a single stage.

In this paper, we investigate the last generation phase (step $3 \rightarrow 4$ in Figure), namely the text generation for the current utterance. In our case, we also distinguish the dialogue act planning substage that consists of selecting the edge type in terms of dialogue acts.

3.2 Linearization

This section describes linearization of graphs representing multi-party dialogues annotated by discourse relations, in addition to the main MPC features. As it was mentioned above, we do not have the goal of tuning the linearization technique, and we have chosen one of the most reasonable ones.

The graph structure can be converted to a sequence by sorting the utterances by time. Each utterance and its meta can be linearized according to the following way. Firstly, we should assign an utterance id and indicate the current speaker. Then we must specify which addressee statement to respond to and how to respond to it (discourse relation). Finally, we should produce the next utterance text. To handle the first two steps, we suggest to use special tokens as the identifiers of speakers and utterances: $\{\langle s_i \rangle\}$ and $\{\langle u_i \rangle\}$ correspondingly. For instance, a linearized i-th utterance written by the j-th speaker in response to the k-th utterance looks like as follows:

"$\langle u_i \rangle$ $\langle s_j \rangle$ $\langle$relation$\rangle$ $\langle u_k \rangle$ response text"

Also, we use a separator token to join single utterances and get full representation of the current dialogue state. To specify an utterance to respond to at the current turn, we add its representation to the end of the dialogue sequence. We use the resulting representations as the inputs of seq2seq models. Figure 3 demonstrates an example of the MPC dialogue linearization procedure.

We passed the target seq2seq texts to the model in the following format: "$\langle$relation$\rangle$ response text".

It should be highlighted that the response text is generated following discourse relations. Consequently, its tokens are produced with an attention mechanism that takes into account the discourse token. Moreover, the transformer-based language modeling approach allows us not to use a special auxiliary model like in (Ji et al., 2016; Harrison et al., 2019).

3.3 Model and Loss Function

We use BART and T5 as the base transformer models due to their state-of-the-art performance in various text generation and graph-to-text generation tasks.

In our approach, discourse tokens are planned first and an auxiliary model is not required, their importance can be adjusted through weights in the loss function. To this end, we employ the weighted cross-entropy loss:

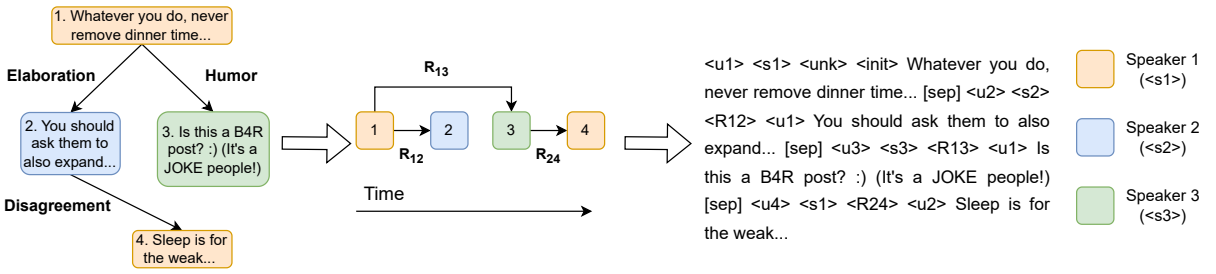

Figure 3: Example of the discursively-annotated MPC linearization process. Firstly, all nodes are ordered temporally, forming a chain. Then, it is transformed to text representation using special tokens to display meta information: $\langle u_i \rangle$ are used for utterance ids, $\langle s_i \rangle$ are tokens for speaker ids (are signified by colors), and $\langle R_{ij} \rangle$ are used for relations. Additionally, an $\langle$init$\rangle$ token is introduced due to the fact that the first replica does not have an adressee.

$$L = -\frac{1}{|S|} \sum_{j=1}^{|S|} \sum_{i=1}^{|D_{\text{all}}|} w(y_j) I\{x_i = y_j\} \log(p(x_{ji})) \tag{1}$$

$$w(y) = \alpha I\{y \in D\} + I\{y \notin D\} \tag{2}$$

Here, $|S|$ is the target sequence length, $|D_{\text{all}}|$ is the full vocabulary size, $p(x_{ji})$ is the predicted probability of the i-th token for the place j, and y_j is the target token. I denotes an indicator function. D is the predefined set of discourse tokens, and α is the weight related to the dialogue acts planning task. When the α coefficient is zero, we actually provide the standard response generation task instead of the multitask learning.

The described approach is quite intuitive, but at the same time, it allows for significant improvement of the quality of generation and acceleration of convergence, as demonstrated in Section 5.

4 Datasets

This section presents the discursively-annotated datasets used for evaluation.

4.1 CDSC

First, we utilize the largest manually annotated dataset of dialogue acts in online discussions, namely the Coarse Discourse Sequence Corpus (CDSC) proposed by Zhang et al. (2017). It contains ~9K Reddit threads (in English), with comments annotated with 9 main discourse act labels that were designed to cover general discourse and an "other" label. It should be highlighted that, to the best of our knowledge, this dataset is the only open-source dataset that is sufficiently large and includes discourse act labeling.

The list of the dialogue acts used in the dataset is the following: "Question", "Answer", "Announcement", "Agreement", "Appreciation", "Disagreement", "Negative Reaction", "Elaboration", "Humor", "Other".

There exists some missing values in the data, and we replace them with the $\langle$unk$\rangle$ special token. We splitted the data into training and test sets with a ratio of 6:1. As a preprocessing, we removed instances with missing values and all non-ascii characters from texts.

4.2 Movie Reddit Dataset

No other large-scale, discursively-annotated open datasets for the MPC generation task are available, and manually-labeled data is not typically accessible in real-world applications. Therefore, we collected our own dataset and labelled it automatically to increase the significance of the findings.

Similar to CDSC, we parsed threads from Reddit (the largest open source of dialogues), but focused primarily on the movie domain since it does not require any specific knowledge and is generally considered for the conversation analysis tasks (Zhou et al., 2018). We collected roughly 90k dialogues from the 25 most popular Reddit subthreads discussing movies, series and TV shows. To obtain discourse acts labels, we trained our own discourse parser from scratch based on the CDSC dataset. Existing parsers are trained using much smaller datasets and operate with other discourse relations, making evaluation inconvenient. We chose the Two-Stage discourse parser (Wang et al., 2017) as the model architecture, since it is open-source and has obtained SOTA results for dialogue discourse parsing. The entire procedure for preprocessing and input data construction used is identical to that of CDSC.

5 Experiments

This section discusses implementation details and experiment results, including human evaluation.

5.1 Implementation Details

We fine-tuned the base-sized BART and T5 models (139M and ∼220M parameters respectively). The maximum source length was set to 1024, and the maximum target length was set to 64 (these values were estimated using the training set). The models were trained on batches of size 2 with a learning rate of 2e-5 during 5 epochs. Other hyperparameters were used by default.

Each model was trained on the GPU Tesla V100 32G for approximately 10 hours.

5.2 Discourse Planning Importance

We use the popular ROUGE-based (Lin, 2004) and BLEU-based (Papineni et al., 2002) scores to automatically estimate the overall generation quality. We calculate it based on the target texts cleared of discourse tokens.

We conducted experiments for the three settings of the dataset used for fine-tuning: (1) $\mathcal{D}$ containing discourse relation tokens in both source and target texts; (2) $\mathcal{D}_1$ only containing discourse relations in source texts; and (3) $\mathcal{D}_2$ having no discourse relations at all (is considered as the baseline model).

We selected the weight α for discourse planning as 100 using grid search for the $\mathcal{D}$ setting. Results further detailed in Section 5.4 indicate that it is better to choose the weight of discourse tokens in the loss function larger than the rest. At the same time, the difference for large values is not significant, so we chose the same value of α for both datasets. This weight was not used for $\mathcal{D}_1$ and $\mathcal{D}_2$ since they do not contain dialogue act tokens in the target texts. Additionally, it should be noted that $\mathcal{D}_1$ can be considered an equivalent for $\mathcal{D}$, where the α coefficient is set to 0 and dialogue acts are not being planned.

Table 1 presents the F1-scores for the ROUGE-based and BLEU-based metrics for the BART model. The results demonstrate that the model incorporating discourse planning (setting $\mathcal{D}$) achieved the highest scores and was significantly superior to the other models. This indicates that discourse planning simplifies the generation of response texts, even for BART.

Furthermore, for the Movie Reddit dataset, the results in setting $\mathcal{D}_1$ outperform those those in setting $\mathcal{D}_2$. It follows that in the case when the training dataset is large and comprises more examples of discourse dependencies, incorporating dialogue act markers in input texts can also be beneficial. Nevertheless, the maximum quality boost is obtained precisely when training the auxiliary discourse planning task (in setting $\mathcal{D}$).

The metrics for $\mathcal{D}$ are slightly lower for the Movie Reddit dataset than for the CDSC. This is primarily attributed to the fact that all dialogue acts in Movie Reddit were labeled automatically, which can lead to inaccurate labels. However, the results remain consistent between the two datasets, and the model featuring a discourse planning stage performs significantly better than the base model, even considering the automatically labeled data.

Table 2 demonstrates results for the T5 model. The language modeling quality is slightly inferior to that of BART, which may be due to a suboptimal hyperparameter selection. The decreased quality in the last rows may be attributed to the use of an extended tokenizer with discourse tokens (nevertheless, this assumption should not greatly affect the quality). At the same time, hyperparameter search was not our primary objective, and the results confirm that with an appropriate selection of hyperparameters, discourse planning greatly improves generation quality.

5.3 Human Evaluation

To enhance the evaluation as well as cover aspects that cannot be assessed by automatic metrics, we conducted a human evaluation. Here, the main goal was to compare texts generated by two BART models: the base model and the model trained via multi-task learning. For each instance, the experts were tasked with choosing which of two options was the best for continuing the dialogue (or whether they were equal), as well as evaluating each option on a 3-point scale according to the criteria of consistency (coherence) and meaningfulness. Coherence assessed the relation between the current utterance and the addressee, as well as the overall logic of the dialogue, while meaningfulness assessed the semantic load of the utterance in its general context. The two scales were rated on a scale of 0-2, with 0 representing a bad prediction, 1 representing a generally normal prediction with some inaccuracies, and 2 representing a prediction close to perfect. In order to ensure reliability in the evaluation, the options were shown in random order.

Dataset	Setting	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2
CDSC	$\mathcal{D}$ [full discourse data]	**9.34**	**0.66**	**8.37**	**8.68**	**0.32**
	$\mathcal{D}_1$ [no discourse in resp.]	7.39	0.48	6.64	6.53	0.30
	$\mathcal{D}_2$ [no discourse at all]	7.44	0.43	6.71	6.58	**0.32**
Reddit	$\mathcal{D}$ [full discourse data]	**8.89**	**0.58**	**7.96**	**8.11**	0.17
	$\mathcal{D}_1$ [no discourse in resp.]	7.76	0.54	7.07	6.80	**0.20**
	$\mathcal{D}_2$ [no discourse at all]	7.45	0.51	6.77	6.47	0.17

Table 1: Performance of BART-based models on the CDSC and Movie Reddit test sets for different variants of training datasets (denoted as settings). We use F1-scores for the ROUGE-based metrics. $\mathcal{D}$ uses discourse relations in both source and target texts in seq2seq training, $\mathcal{D}_1$ has responses cleared of discourse relations, and $\mathcal{D}_2$ is the dataset without discourse relations at all (is used to train the baseline model). Here, STD ≤ 0.6 in cases of unigram-based metrics and STD ≤ 0.1 in cases of bigram-based metrics.

Setting	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2
$\mathcal{D}$ [full discourse data]	**8.81**	**0.50**	**7.87**	**8.02**	0.25
$\mathcal{D}_1$ [no discourse in resp.]	7.06	0.39	6.36	6.12	0.25
$\mathcal{D}_2$ [no discourse at all]	6.94	0.41	6.24	5.96	0.21

Table 2: T5-based model performance on the CDSC test set for different training datasets and α coefficients.

Model	# better	Coherence	Meaning.
Base	62	1.11	1.32
Disco	**83**	**1.32**	**1.33**

Table 3: Human evaluation results on the subset of 200 dialogues from the CDSC test set. "Base" refers to the base BART model and "Disco" refers to the model trained via dialogue acts planning.

Table 3 presents the obtained results for 200 random dialogues from the CDSC test dataset. Here, the scores are averaged across the corpus. We can see that responses produced by the custom approach are preferable in more cases. This is mainly because the discourse-based model's responses are more coherent and more appropriate for continuing the dialogue, despite perhaps less semantically appropriate formulations (the task of generating texts for some dialogue acts is quite challenging). Although the overall improvement is not substantial, there is a considerable progress in the aspect of consistent dialogue generation.

5.4 Convergence Speed

In this section, we evaluate the convergence speed of our model. The rate of convergence can be estimated in several ways, and in this case we have chosen one of them, which is related to estimating the fewest number of steps to get high quality. For early quality estimation, we train models for a smaller number of steps (2 epochs) with varying values of the α coefficient in the loss function to indicate the importance of the discourse planning task. These values are 1, 10, 30, 100 and 200. We measure the quality of discourse tokens using Accuracy and the quality of response texts using F1-based ROUGE-L.

Figure 4 demonstrates the results that reveal a strong correlation between Accuracy and ROUGE values, suggesting that improved discourse planning improves the overall quality of language modeling. Furthermore, these results indicate that the approach converges faster (reaching optimal quality at earlier epochs) if the discourse planning task has more weight. For instance, increasing α from 1 to 100 yields a significant increase in the convergence speed of the training process, requiring far fewer steps to attain the best possible generation quality.

6 Discussion

In this section, we partially explain how discourse improves generation quality for the model trained using discourse planning and demonstrate differences using concrete examples.

6.1 Error Analysis

In order to analyze which dialogue acts the discourse-based model actually plans and which of them can improve the overall quality of language modeling, we compare the quality of dialogue acts planned by the base and the discourse-based BART

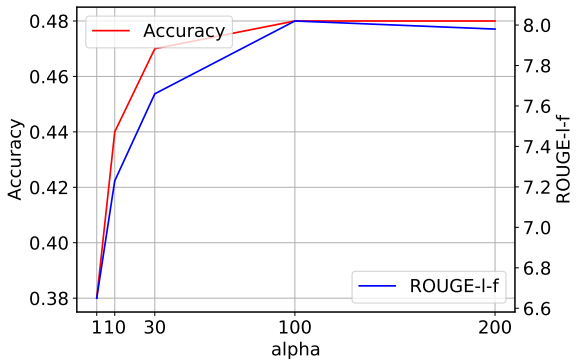

Figure 4: Discourse Accuracy (blue line) and ROUGE (red line) scores depending on α for the CDSC test set.

models. As the base model does not explicitly generate dialogue acts, our trained discourse parser was used to label them. The corpus used for training also contains unlabeled instances, which are not taken into account in our analysis.

Figure 5 demonstrates the confusion matrices for both the base and custom models. The correct labels were taken from the dataset. The results illustrate that the base model achieves only an accuracy of 0.315, whereas the custom model gets 0.615. The task is complicated by the fact that the dialogue can be continued in various ways and often there is no single correct dialogue act.

The confusion matrix for the discourse model is closer to diagonal, indicating improved performance. A standout feature is that the custom model successfully plans not only common relations, such as Elaboration and Answer, but also rarer ones. So, the discourse-based model more accurately determines when it is necessary to thank the interlocutor (Appreciation), and when to ask a clarifying question (Question). Interestingly, the discourse model predicts better even such relations as Disagreement and Humor. Some relations are quite non-trivial, and even with the right relations planned, it can be challenging for the model to generate the correct words to achieve the highest automatic generation metrics. However, as seen in Section 5, the right choice of dialogue acts is a step towards high-quality generation.

6.2 Generation Examples

Figure 6 shows examples of dialogues generated by the base BART model and the BART model fine-tuned using discourse tokens. We focus on the generation of the last utterance, as this is the second stage of the general generative MPC pipeline.

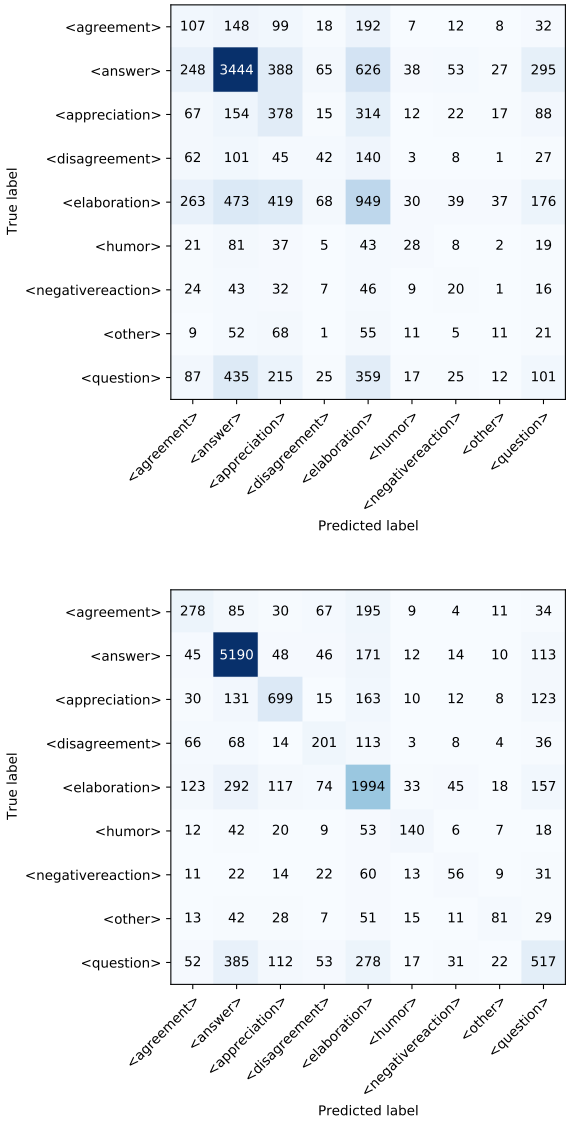

Figure 5: Confusion matrices of dialogue act planning for the CDSC test set for the base BART (top) and discourse BART (bottom) models.

The first example demonstrates a simple dialogue with only two speakers. Even in such scenarios, the base BART model may struggle. In this instance, the base model attempted to answer the question “maybe an endgame companion?”, while completely disregarding the context of the conversation. At the same time, the discourse planning model was able to respond in a more logical and reasonable manner, continuing the topic and aiming to achieve the initial goals of the first speaker by asking a new question.

The second dialogue presented appears to be a chain, with three speakers. One can see that the most pertinent relations for this dialogue are “agree-

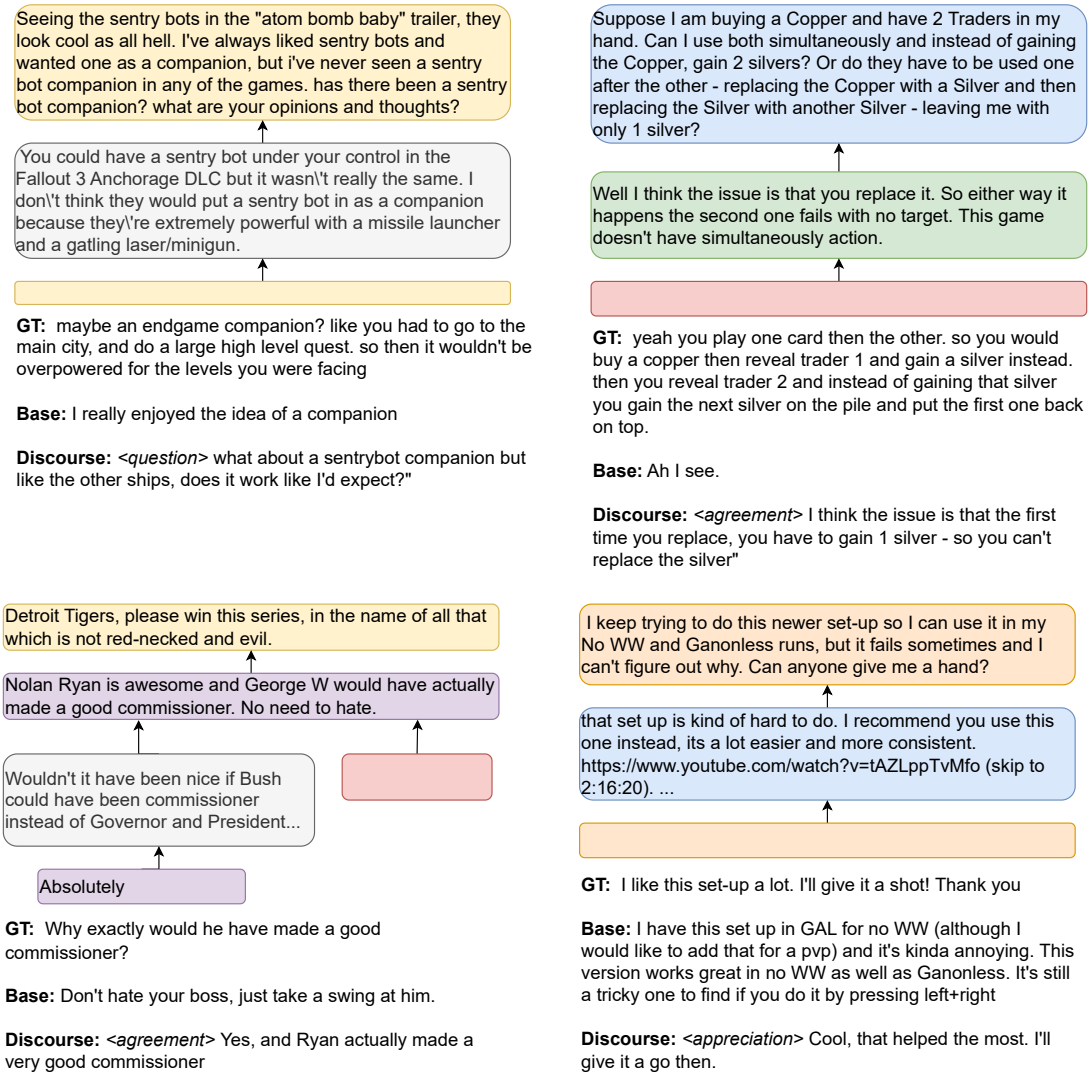

Figure 6: Examples of generated dialogues with the ground-truth reply (GT) and options generated by the base and discourse approaches. Node color is used to identify the speaker and arrows on the edges indicate the addressees. The task is to generate the utterance text for the empty node. Here, "Base" refers to the base BART model and "Discourse" refers to the BART model fine-tuned using special discourse tokens. The discourse-based model also produces discourse tokens which are shown in italics.

ment" and "disagreement", given the introduction of a new speaker. Nevertheless, the base model chooses options with "appreciation", which does not contribute to dialogue continuation.

The third example (bottom left in the Figure) demonstrates a dialogue with a more intricate structure, making the planning of discourse relations more challenging. There exist several ways to move the dialogue forward, and the chosen "agreement" relation allows the discourse-based model to generate an utterance that is not removed from the context.

The final example shows the case in which the "appreciation" token is sufficient for the discourse-based model to generate a concise, suitable answer without excessive detail.

It is important to emphasize that the base model is still capable of producing valid responses due to its element of randomization in the sampling process. Nevertheless, the accurate choice of a discourse relation (dialogue act) at the start significantly simplifies the search for alternatives and almost always results in valid responses.

7 Conclusion and Future Work

In this paper, we explored the effectiveness of transformer fine-tuning based on discourse dialogue acts planning for the multi-party conversation generation task. We evaluated our approach on the largest manually and automatically annotated datasets of

dialogues from Reddit.

The evaluation including automatic and human assessments revealed that incorporating special discourse tokens into the linearized training sequences could significantly improve the generation metrics and is an important step towards coherent generation. The proposed approach performed well even on the automatically annotated dataset, and increasing the weight of discourse tokens in the loss function further accelerated learning convergence.

Future work includes the analysis of other types of linguistic information (such as syntactic and semantic relations), other ways of integrating them into the training process, as well as experiments for alternative MPC generation pipelines.

8 Ethics and Broader Impact

The training of large transformer-based models is one of the reasons leading to global warming. However, we do not train these models from scratch and use the fine-tuning procedure. Moreover, we consider only the base variants of the models that have a lower number of trainable parameters.

9 Limitations

The proposed approach is not limited to the English language or BART/T5 approaches. The main limitations are the presence of annotated data that can be acquired manually or with the help of a parser, and the seq2seq nature of the transformer-based approach. As with most conversational agents, there are possible adverse impacts, like spreading harmful or hateful messages or misinformation. Models mainly learn the training dialogues, and most of these issues can be addressed through proper preprocessing or the selection of appropriate datasets.

Acknowledgements

The article was prepared within the framework of the HSE University Basic Research Program. It was also supported in part through the computational resources of HPC facilities at NRU HSE.

References

Stergos D. Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 928–937. The Association for Computational Linguistics.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Alexander Chernyavskiy. 2022. Improving text generation via neural discourse planning. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 1543–1544. ACM.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3808–3814. ijcai.org.

Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019. Interactive matching network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2321–2324. ACM.

Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3682–3692. Association for Computational Linguistics.

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn A. Walker. 2019. Maximizing stylistic control and semantic accuracy in NLG: personality variation and discourse contrast. *CoRR*, abs/1907.09527.

Alexander Miserlis Hoyle, Ana Marasovic, and Noah A. Smith. 2021. Promoting graph awareness in linearized graph-to-text generation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 944–956. Association for Computational Linguistics.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. GSN: A graph-structured network for multi-party dialogues. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5010–5016. ijcai.org.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. *CoRR*, abs/1603.01913.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 97–102. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Top-down discourse parsing via sequence labelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 715–726. Association for Computational Linguistics.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1909–1919. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Zihao Zheng, Heng Zhang, Bing Qin, Min-Yen Kan, and Ting Liu. 2021a. Dadgraph: A discourse-aware dialogue graph neural network for multiparty dialogue machine reading comprehension. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE.

Shimin Li, Hang Yan, and Xipeng Qiu. 2021b. Contrast and generation make BART a good dialogue emotion recognizer. *CoRR*, abs/2112.11202.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Yi Luan, Yangfeng Ji, and Mari Ostendorf. 2016. LSTM based conversation models. *CoRR*, abs/1603.09457.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2133–2143. The Association for Computational Linguistics.

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 of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. *CoRR*, abs/2007.08426.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1577–1586. The Association for Computer Linguistics.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7007–7014. AAAI Press.

Matthew Stone, Una Stojnic, and Ernest Lepore. 2013. Situated utterances and discourse relations. In *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013, March 19-22, 2013, University of Potsdam, Potsdam, Germany*, pages 390–396. The Association for Computer Linguistics.

David R. Traum. 2003. Issues in multiparty dialogues. In *Advances in Agent Communication, International Workshop on Agent Communication Languages, ACL 2003, Melbourne, Australia, July 14, 2003*, volume 2922 of *Lecture Notes in Computer Science*, pages 201–211. Springer.

David C. Uthus and David W. Aha. 2013. Multiparticipant chat analysis: A survey. *Artif. Intell.*, 199-200:106–121.

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 2021. A structure self-aware model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3943–3949. ijcai.org.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-based deep matching of short texts. *ArXiv*, abs/1503.02427.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 184–188. Association for Computational Linguistics.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4970–4977. AAAI Press.

Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Eleventh International AAAI Conference on Web and Social Media*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 708–713. Association for Computational Linguistics.