GroundHog: Dialogue Generation using Multi-Grained Linguistic Input

Alexander Chernyavskiy[1], **Lidiia Ostyakova**[1,2], **and Dmitry Ilvovsky**[1]
[1] HSE University, Russia
[2] Moscow Institute of Physics and Technology, Russia
alschernyavskiy@gmail.com, ostyakova.ln@gmail.com
dilvovsky@hse.ru

Abstract

Recent language models have significantly boosted conversational AI by enabling fast and cost-effective response generation in dialogue systems. However, dialogue systems based on neural generative approaches often lack truthfulness, reliability, and the ability to analyze the dialogue flow needed for smooth and consistent conversations with users. To address these issues, we introduce GroundHog, a modified BART architecture, to capture long multi-grained inputs gathered from various factual and linguistic sources, such as Abstract Meaning Representation, discourse relations, sentiment, and grounding information. For experiments, we present an automatically collected dataset from Reddit that includes multi-party conversations devoted to movies and TV series. The evaluation encompasses both automatic evaluation metrics and human evaluation. The obtained results demonstrate that using several linguistic inputs has the potential to enhance dialogue consistency, meaningfulness, and overall generation quality, even for automatically annotated data. We also provide an analysis that highlights the importance of individual linguistic features in interpreting the observed enhancements.

1 Introduction

Text generation methods, particularly for conversational systems, have become increasingly popular in recent years. The conversational systems play a crucial role in enhancing the effectiveness of user-agent interactions (Young et al., 2018; Gu et al., 2019; Le et al., 2019). Dialogue systems are used for human-machine conversations on various topics. Some systems are built as question-answering systems or personal assistants, focusing on specific domains or general inquiries.

Despite showing impressive response generation capabilities, language models, even ones like GPT-4, have shortcomings in terms of truthfulness (OpenAI, 2023). Consequently, researchers are exploring methods to combine generative and extractive approaches in order to make the responses of dialogue systems more logical and reliable. Here, the primary objective is to incorporate external knowledge, resources, or databases into the response generation process. The previous studies have demonstrated a substantial enhancement in the quality of generation by incorporating grounding, which improves the factual accuracy of the responses (Feng et al., 2020). Grounding input is commonly integrated into dialogue generation models along with the context of a particular utterance (Zhao et al., 2020) or a preceding part of the dialogue that represents the conversational history (Rashkin et al., 2021).

Furthermore, previous works have explored leveraging grounding in combination with other features, including commonsense and named entities (Varshney et al., 2022; Wu et al., 2022), dialogue acts (Hedayatnia et al., 2020), topic shifts (Wu and Zhou, 2021), discourse annotation (Khalid et al., 2020), to improve dialogue generation. Despite the fact that additional linguistic features are frequently used to improve the consistency of generated dialogues (Ji et al., 2016; Harrison et al., 2019), previous studies focused on individual and superficial examination of linguistic features. In our research, we conducted a more comprehensive analysis, evaluating the relative significance of each of them and the overall contribution.

We primarily investigate the impact of various linguistic features on response generation in a multi-grained input framework. Specifically, we analyze the effects of semantic relations derived from Abstract Meaning Representation (AMR) (Banarescu et al., 2013), dialogue acts extracted from dialogue discourse trees (Stone et al., 2013; Zhang et al., 2017), and utterance-based sentiment representation.

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 149–160

Experiments on response generation are generally conducted using open-source sequence-to-sequence models (Raffel et al., 2020; Rashkin et al., 2021). Among these models, the BART architecture (Lewis et al., 2020) has gained significant popularity due to its state-of-the-art performance in various text generation tasks. Due to its efficiency in processing linearized inputs, it is often utilized in graph2text tasks (Ribeiro et al., 2020). Moreover, this capability can be further extended to analyze conversation graphs. However, the length of input texts can often present challenges for Transformer-based models. In this study, we introduce GroundHog, an approach that uses multiple input encoders to preserve input information effectively.

Our contributions can be summarized as follows:

- We present a novel dataset consisting of open-domain conversations for dialogue system training. This dataset is augmented with linguistic features and grounding, enhancing its potential for training high-quality models.
- We propose the use of grounding and linguistic features for response generation in dialogue systems. An ablation study is conducted to analyze their individual contributions.
- A modification to the BART architecture is suggested to effectively capture long multi-grained inputs.
- We perform an analysis to interpret the improvements and discuss our findings.

2 Dataset

The most popular datasets, including open-domain conversations grounded in Wikipedia information, are *Wizard of Wikipedia* (Dinan et al., 2018) and *CMU DoG* (Zhou et al., 2018). To narrow the scope of this study and facilitate the language model training, CMU DoG was used as a starting point. This dataset contains 4112 grounded conversations devoted to the discussion of Wikipedia articles about popular movies. To extend the dataset, we collected Reddit[1] conversations on the same topic in English. Specifically, we parsed conversations from the 25 most popular subreddits related to films, series, and TV shows. These subreddits provided discussions that were tied to specific topics or comments. Additionally, we gathered comments that mentioned key phrases such as “movie” and “film”.

[1] https://www.reddit.com/

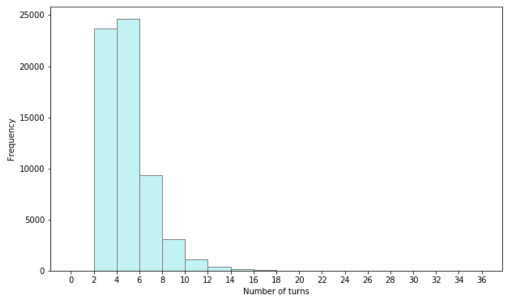

Figure 1: Distribution of dialogue lengths in collected dataset

The dataset preprocessing stage involved removing images, extra symbols, and emojis, as these were not considered in our research. In total, our collected dataset consists of approximately 62,500 multi-party dialogues, with an average of 5 turns per conversation (see Figure1). The length of extracted Reddit conversations is significantly shorter compared to CMU DoG dialogues, which have an average of 21.43 turns per conversation.

The dataset contains conversations collected along with linguistic annotations, grounding, and meta information related to each extracted dialogue. Specifically, automatically retrieved linguistic features for each turn in the dataset are presented in the following format:

- discourse annotation is represented as identifiers of connected turns with a discourse class describing the relation between them;
- sentiment class of a turn, accompanied by its probability;
- AMR graph is provided in simplified form for each turn.

The process of annotating data is described in detail in Section 3.1. Our final dataset is publicly available at the link: https://huggingface.co/datasets/alexchern5757/groundhog_reddit.

It should be emphasized that all datasets containing open-domain dialogues share the same limitations related to grounding. Casual conversations are distinguished by the absence of rigid topic boundaries, stylistic ambiguity, and a strong reliance on context. Evaluative information in these dialogues is often presented as facts, which can result in inaccurate grounding extraction.

3 Methods

3.1 Dialogue Features

In order to generate coherent and truthful responses, we incorporate grounding and several linguistic features that describe the current dialogue state as model inputs.

Discourse Discourse can be represented in various ways, with one of the most widely used approaches being Rhetorical Structure Theory (RST) for plain texts (Mann and Thompson, 1988). RST employs elementary discourse units to analyze the structure of the text, whereas in dialogue analysis, trees are constructed over utterances. Dialogue discourse graphs, as introduced by Stone et al. (2013), extend the concept of standard dialogue graphs by including discourse labels for each utterance indicating the specific function or pragmatic purpose of the utterance (e.g., Disagreement, Appreciation, Question). An example is provided in Appendix A.

The application of discourse annotation, combined with grounding techniques, has demonstrated the potential for generating media dialogues that are more consistent and truthful (Majumder et al., 2020; Chernyavskiy and Ilvovsky, 2023). This integration of linguistic features and grounding methods has shown promise in enhancing the quality of such tasks.

In order to achieve automatic discourse annotation, we implemented and trained the parser model suggested by Shi and Huang (2019). The training process was started from scratch and utilized the Coarse Discourse Sequence Corpus (CDSC) (Zhang et al., 2017), which is the largest manually annotated dataset of discourse acts in online discussions.

Abstract Meaning Representation (AMR) Abstract Meaning Representation is based on directed acyclic graphs and provides a structured semantic representation of language, including semantic role annotations consisting of arguments and values (Banarescu et al., 2013). Given that incorporating AMR graphs enhances task-oriented dialogue generation (Yang et al., 2023) and the promising prospects of integrating AMR with pragmatic intents (Bonial et al., 2020), we use these graphs as one of the linguistically motivated inputs in the experiments.

In our dataset, an AMR graph was generated for each sentence within an utterance, and then these subgraphs were combined into a single graph. To reduce the complexity of the representation, we truncated vertices at a depth beyond a specified constant. A more detailed description of the AMR graphs is provided in Appendix A. We adopt a similar method to linearize AMR graphs, as proposed by Ribeiro et al. (2020).

Sentiment The sentiment labels assigned to each utterance in the dataset indicate the polarity of the sentiment expressed, using a 3-point scale: Positive, Negative, or Neutral. The RoBERTa model, which was trained on tweets, was utilized for the corresponding labeling task (Barbieri et al., 2020). To incorporate information about sentiment, special tokens were integrated into linearized representations of the dialogues.

Grounding Grounding is an important aspect of model input as it serves to mitigate the issues associated with hallucinations in language models. Generally, when the utterance does not pertain to an opinion, the main fact can be derived from the provided grounding.

There are several approaches to fact-control realization for overcoming hallucinations within a dialogue system. One of them is the use of external memory, which was proposed in RETRO (Borgeaud et al., 2022) and KELM (Lu et al., 2021) models when the relevant parts of the training texts are passed to the cross-attention mechanism at the stage of next response generation. An alternative method is to extract grounding text from external databases, for instance, by using web mining like in the Sparrow (Glaese et al., 2022) approach. LaMDA (Thoppilan et al., 2022) proposes an approach combining structured factual grounding from an external knowledge base (Google Search API) and dialogue context both in the training and inference stages.

In this paper, we focus on the Sparrow approach and explore the importance of using grounding for generating consistent open-domain dialogues. We use the MediaWiki API[2] to conduct searches for two types of queries: movie titles and entire Reddit thread titles. A restriction was imposed to retrieve a maximum of five documents for each query. Subsequently, a summarized version of these documents was created, consisting of five sentences. These summaries were then combined into a single grounding text.

[2] https://github.com/goldsmith/Wikipedia

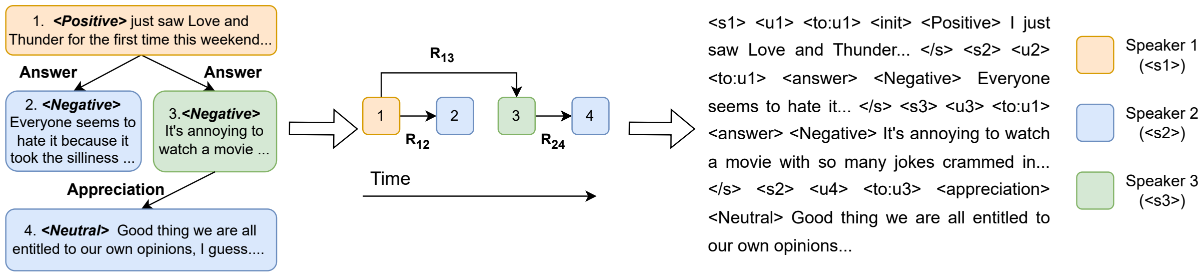

Figure 2: Example of the discursively annotated conversation linearization process. Firstly, all nodes are ordered temporally, forming a chain. Then, it is transformed into text representation using special tokens to display meta information: $\langle u_i \rangle$ are used for utterance ids; $\langle s_i \rangle$ are tokens for speaker ids (are signified by colors); $\langle \text{to:}u_i \rangle$ are used for addressees; and $\langle R_{ij} \rangle$ are used for relations. Additionally, an $\langle \text{init} \rangle$ token is introduced due to the fact that the first replica does not have an addressee.

3.2 Dialogue Linearization

The linearization of dialogue graphs plays a crucial role in our approach. Hoyle et al. (2021) demonstrated that Transformers exhibit invariance to the specific method employed for linearization. Therefore, we employ discourse and AMR graphs for dialogue modeling, followed by a thoughtful linearization process.

Our linearization procedure is implemented in the following way. Firstly, all utterances are arranged in chronological order to establish a linear sequence. Secondly, each utterance is linearized independently, taking into account its own characteristics as well as the attributes of the connecting edge to its addressee. To achieve this, each utterance is assigned a unique identifier, the current speaker is indicated, and the addressee statement to which the utterance responds is specified. Thirdly, the appropriate response strategy is determined, as indicated by a discourse relation and sentiment tokens. Finally, the text of the subsequent utterance is incorporated. We utilize special tokens to identify speakers, utterances, and addressees, namely $\{\langle s_i \rangle\}$, $\{\langle u_i \rangle\}$ and $\{\langle \text{to:}u_i \rangle\}$ respectively. As an example, a linearized i-th utterance written by the j-th speaker in response to the k-th utterance has the following form: "$\langle s_j \rangle$ $\langle u_i \rangle$ $\langle \text{to:}u_k \rangle$ $\langle \text{relation} \rangle$ $\langle \text{sentim.} \rangle$ text".

We employ a separation token to combine individual utterances and create a full representation of the dialogue state. Figure 2 provides an example of the conversation linearization procedure. By eliminating all text and sentiment tokens, the linearized representation can be conveniently converted into a raw linear discourse representation.

3.3 GroundHog Model

We suggest the GroundHog model as an effective neural approach for encoding diverse types of input information. It incorporates multiple Transformer-based encoders to capture multiple levels of granularity in the input data. Unlike previous approaches such as Longformers (Beltagy et al., 2020), our focus is on the attention mechanism within each input rather than utilizing global attention. In addition, we reduce the size of the attention matrices compared to Longformers.

The architecture is based on the customized BART, as illustrated in Figure 3. Our approach involves the utilization of multiple texts as input, on which it does not formally impose restrictions. The first input text should contain the primary information, whereas the others should provide supplementary information. In our case, the inputs are the following: (1) a dialogue history that has been enriched with discourse and sentiment tokens; (2) a raw, linearized representation of a discourse dialogue graph; and (3) an addressee's utterance and a part of its AMR graph.

Each input is first processed through a common tokenizer and then encoded separately using its own BART encoder. In order to create a more universal approach, embeddings from *all* inputs could be aggregated through convolution. However, this would substantially change the standard input format of the pre-trained BART decoder, making the training process more challenging without a large dataset for additional pre-training. Therefore, we divide the inputs into two categories: the main text and the supplementary texts.

The model does not modify the embedding of the main text before the decoder, and it retains the attention mask for this text. The other inputs are

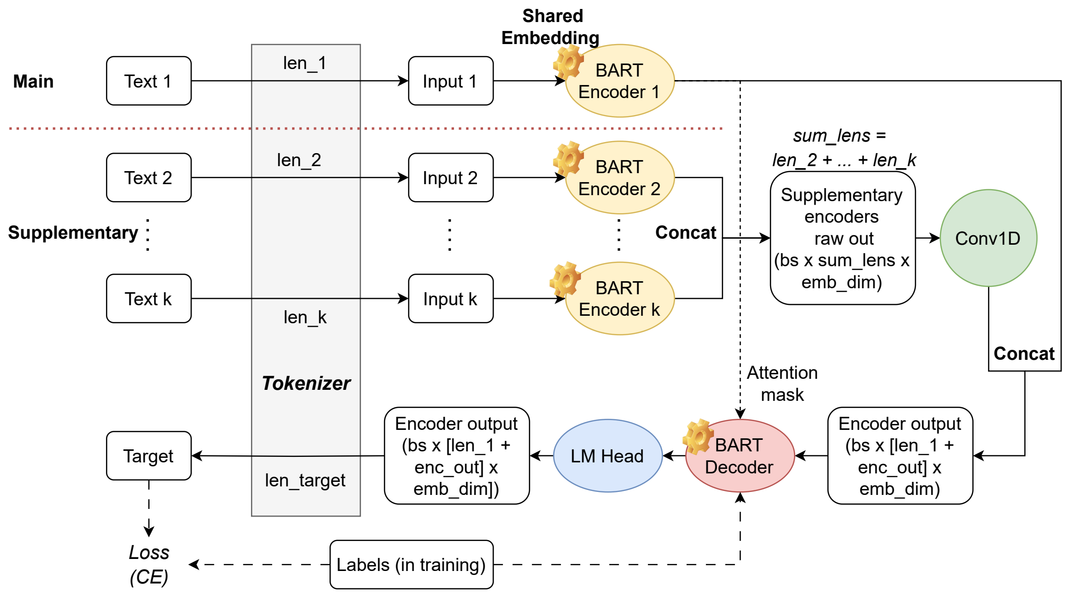

Figure 3: GroundHog architecture. The GroundHog architecture comprises individual BART encoders for each input text, which are subsequently aggregated and used as input for the BART decoder. To reduce the dimensionality of the inputs, a 1D convolutional layer is applied to all inputs except the main input. The shared embedding layer is denoted by the gear icon. In addition, intermediate tensor dimensions are indicated (batchsize is denoted as bs).

combined using concatenation and a convolutional layer. However, this approach may introduce some disruption to the token order, and consequently, the attention masks from the encoder for these inputs are not utilized in the decoder input. In this research, we conducted experiments using different aggregation methods and determined that the one-dimensional convolutional layer yielded the most favorable results.

As in the base model, the language modeling head is utilized after the decoder. We use the same tokenizers and shared embedding layer for all encoders and the decoder. As is common in language modeling decoder-based approaches, we employ a standard cross-entropy loss.

4 Experiments

4.1 Implementation Details

We fine-tuned the base-sized BART (139M parameters) model and the GroundHog models based on it. We used various lengths for different inputs but the maximum was 1024 tokens. The models were trained on batches of size 2, with a learning rate of 2e-5, for 5 epochs. For all other hyper-parameters, we used the default values.

All parsers and datasets used have the open source MIT license.

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

4.2 Automatic Evaluation

In order to conduct a more comprehensive analysis of the generation of complex responses, we divided the dataset into two subsets: dialogues with long last responses (consisting of at least two sentences) and dialogues with short responses.

We conducted experiments using both the BART and GroundHog models for the several configurations of the dataset used for fine-tuning:

- In $\mathcal{B}$, we fine-tuned the base BART model using the concatenation of the dialogue history, thread title, and grounding as the input.
- In $\mathcal{G}_1$, we trained the GroundHog model using the concatenation of the dialogue histories and thread titles.
- In $\mathcal{G}_2$, we extended input from $\mathcal{G}_1$ by adding grounding.
- In $\mathcal{G}_3$, we enriched the dialogue history from $\mathcal{G}_2$ by discourse linguistic tokens.
- In $\mathcal{G}_4$, we added separate linguistic inputs associated with AMR: (1) AMR for the full dialogue history (concatenated representations of single utterances); (2) AMR for the addressee.
- In $\mathcal{G}_5$, we extended the input from $\mathcal{G}_4$ by adding sentiment tokens.

In all cases where grounding was utilized, it was concatenated with the main text input. This was necessary to ensure that the attention mechanism

Model	Setting	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2
BART	$\mathcal{B}$ [history; title; grounding]	17.71	3.69	15.95	17.2	2.86
GroundHog	$\mathcal{G}_1$ [history; title]	17.79	3.71	16.05	16.99	2.88
	$\mathcal{G}_2$ [+grounding]	17.86	3.85	16.08	**17.32**	3.00
	$\mathcal{G}_3$ [+discourse]	17.88	3.87	16.09	17.15	3.04
	$\mathcal{G}_4$ [+AMR]	17.88	3.80	16.17	17.26	2.94
	$\mathcal{G}_5$ [+sentiment]	**17.91**	**3.93**	**16.19**	17.25	**3.09**

Table 1: Model performance on the test set (**long responses**) for different model input settings. $\mathcal{B}_i$ and $\mathcal{G}_i$ are related to the BART and GroundHog models trained using different combinations of inputs. Here, the standard deviation is less than 0.007 in all cases.

Model	Setting	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2
BART	$\mathcal{B}$ [history; title; grounding]	9.86	2.10	8.98	9.18	1.65
GroundHog	$\mathcal{G}_1$ [history; title]	9.66	1.88	8.77	8.90	1.40
	$\mathcal{G}_2$ [+grounding]	9.85	2.16	8.98	9.20	1.70
	$\mathcal{G}_3$ [+discourse]	10.11	2.37	9.21	9.46	1.90
	$\mathcal{G}_4$ [+AMR]	9.82	2.06	8.93	9.12	1.56
	$\mathcal{G}_5$ [+sentiment]	**10.23**	**2.47**	**9.32**	**9.52**	**1.98**

Table 2: Model performance on the test set (**short responses**) for different model input settings.

adequately considered the specific components of grounding. Simultaneously, grounding was treated as a distinct input due to its voluminous nature, which may necessitate its truncation.

For automatic evaluation of generated responses, we calculated the ROUGE-based[3] (Lin, 2004) and BLEU-based[4] (Papineni et al., 2002) scores using target texts cleared of special tokens (raw texts). The obtained scores (mean F1 over three runs) are presented in Table 1 for the long texts.

The GroundHog model ($\mathcal{G}_2$) exhibited superior performance compared to BART across all metrics when provided with the same inputs. This suggests that longer inputs are more effectively processed when handled separately. However, it is important to note that one limitation of the GroundHog model is that its decoder requires substantial amounts of training data to learn effectively from scratch. With sufficient pretraning, these results can be improved. Also, triggered by this limitation, we conducted a grid search and determined that setting the embedding size after the 1D convolutional layer in GroundHog to 256 would prevent an unnecessary increase in the decoder's hidden state.

The results demonstrate that grounding has a positive impact on the ROUGE and BLEU scores. This can be attributed to the fact that the generated responses exhibit a higher level of accuracy in terms of factual information. However, the observed difference is not statistically significant. It can be attributed to the subjective nature of most of the generated responses, which often involve personal evaluations. Additionally, the quality of the extracted grounding is not optimal, as it is extracted based on the full dialogues.

It was observed that the inclusion of linguistic features in the model led to improved performance compared to a model trained without these features. Specifically, the model that utilized all linguistic inputs ($\mathcal{G}_5$) achieved the highest scores. The incremental addition of linguistic features resulted in a monotonic improvement in the ROUGE-L metric as well as an overall improvement in other scores.

Furthermore, the experiments indicated that the use of AMR may not be as promising as dialogue acts or sentiment. This finding can be attributed to the complexity of the AMR structure.

Results for the short target responses are presented in Table 2. The overall conclusions drawn from the analysis are consistent with the findings discussed earlier, with the exception of the monotonous growth of metrics. However, a noteworthy observation is that even for short responses, the GroundHog model exhibits superior performance compared to the base BART model.

4.3 Human Evaluation

A human evaluation was conducted to enhance the assessment process and account for aspects that cannot be evaluated using automated metrics. The main objective of this evaluation was to compare

[3] https://pypi.org/project/rouge/

[4] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

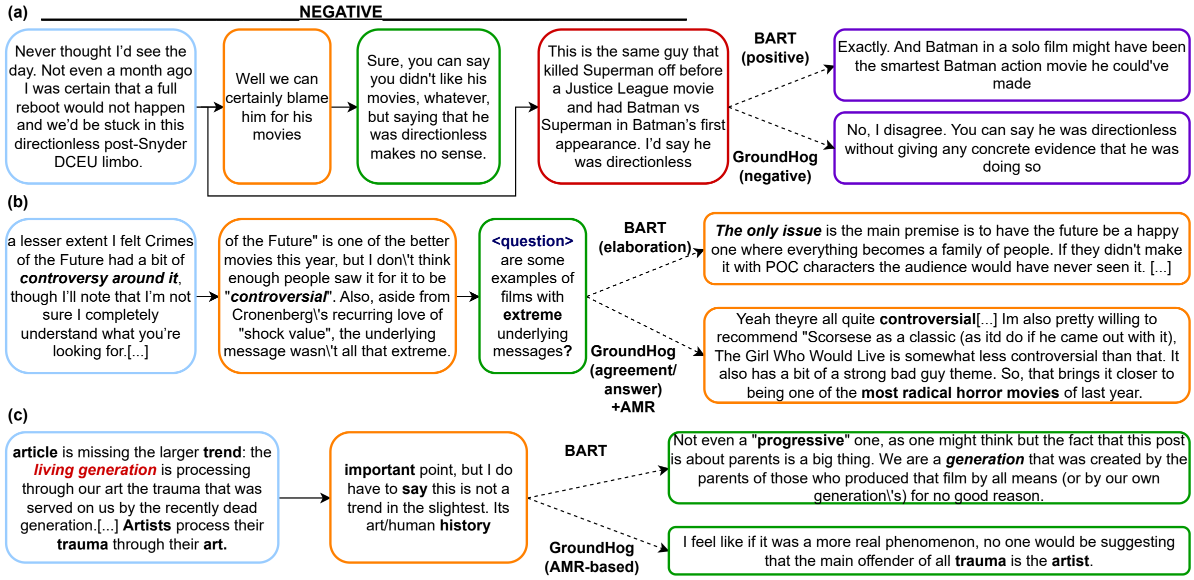

Figure 4: Examples of response generation by the base BART model and the GroundHog model fine-tuned with linguistic inputs. Each color represents a different speaker. The task was to generate text in the last utterance.

the texts generated by the BART model ($\mathcal{B}$) with those of the GroundHog model employing linguistic features ($\mathcal{G}_5$). Experts were tasked with determining the preferable option for continuing the conversation, or whether the alternatives were equal. Also, each option was evaluated on a 3-point scale based on coherence (utterance-based), meaningfulness, and consistency (dialogue-based) criteria.

Dialogue consistency assessed the connection between the current utterance and the addressee, as well as the overall logical progression of the dialogue. Meaningfulness assessed the semantic load of the utterance within its general context. Utterance-based coherence was assessed by evaluating the internal coherence of the utterance.

To ensure reliability in the evaluation process, the three scales were rated on a scale ranging from 0 to 2 (0 for poor prediction, 2 for good prediction). To minimize any potential bias, the options for rating were presented in a random order.

Table 3 presents the evaluation results obtained from 250 randomly selected dialogues from the test dataset. The linguistic approach, as observed, generates responses that are preferred in a larger number of cases. Additionally, these responses are more coherent, suitable for continuing the conversation, and formulated with better semantic appropriateness. While the overall improvement is not sizeable, there is notable progress in the generation of consistent conversations.

	# better	Coherence	Meaning.	Consist.
$\mathcal{B}$	79	1.48	1.27	1.38
$\mathcal{G}_5$	**101**	**1.53**	**1.38**	**1.40**

Table 3: Human evaluation results on the random test subset of 250 dialogues.

5 Discussion

In this section, our major objective is to gain a deeper understanding of the linguistic features that contribute to the improved quality of GroundHog. To this end, we conduct a comparative analysis of the texts generated by BART ($\mathcal{B}$) and the texts generated by GroundHog ($\mathcal{G}_5$).

Regarding the interpretation of grounding, its incorporation enhances the factual component of generation. However, the qualitative aspects of grounding in our dataset are not very robust, and it can be a direction for further research.

Sentiment We started our investigation with the analysis of sentiment due to its ease of interpretation. To assess the sentiment in the generated texts, we utilized the same classifier that was applied to the training dataset. The results yielded an overall accuracy of 0.43 for the BART model and 0.44 for the GroundHog model, with no sizeable difference observed. It is worth noting that the majority of texts in the dataset were negative or neutral, as users generally tend to criticize films or actors. Specifically, there were 1525 negative utterances, 1374 neutral utterances, and 841 positive ut-

Model	answer	elaboration	agreement	other	disagreement	appreciation	question	negative	humor
$\mathcal{B}$	1491	**1231**	446	211	158	107	80	15	1
$\mathcal{G}_5$	1508	1174	**454**	223	159	101	**97**	**22**	2

Table 4: Statistics of dialogue acts in texts generated by the base BART and GroundHog models.

terances within the test dataset. Consequently, the focus should be shifted to generating more accurate negative responses. In terms of these responses, the base BART model achieved an F1-macro score of 0.461, while the GroundHog model achieved an F1-macro score of 0.487. This improvement is particularly noteworthy as it leads to an overall enhancement in language modeling.

Figure 4 (a) presents an illustrative MPC example. It is observed that all input utterances within the dialogue are negative in nature. Consequently, the subsequent utterance should also embody a negative sentiment, either by aligning with the general criticism of the film director or by critiquing the statements expressed by other participants. In this context, it can be inferred that GroundHog has produced an appropriate response. Conversely, the response generated by the base BART model is positive in sentiment and considered inappropriate.

Discourse Dialogue acts contribute to dialogue-based consistency and, to some extent, utterance-based coherence. Since existing models do not explicitly generate dialogue acts, we utilized a trained discourse parser to label these acts for comparison with the original responses. The GroundHog model had a higher accuracy score of 0.551 compared to 0.538 for the base model. The confusion matrices showed similar patterns, but there was a slight difference in the distribution of dialogue acts (see Table 4). Specifically, the base model exhibited a higher frequency of "Elaboration", while the GroundHog model generated less common relations such as "Question" and "Agreement". This indicates that the linguistic model's responses are more diverse without compromising their quality.

In the conversation depicted in Figure 4 (b), it can be observed that the custom model response exhibits better consistency. The most correct target response should include the Answer or Agreement relations rather than Elaboration. Unlike the BART model, which lacks information about the previous response being a question, the GroundHog model incorporates this knowledge in order to generate a response that is discursively consistent. Moreover, GroundHog aims to incorporate the main AMR entities, such as the concept of "controversial."

AMR Interpreting the impact of AMR representations is challenging due to their inherent complexity. Generally, AMR has a direct influence on the semantic aspect, specifically the representation of entities and their relations. In this regard, human evaluation has shown that the scores for the criterion of "meaningfulness" are higher for GroundHog texts compared to BART texts.

Figure 4 (c) provides a concrete example illustrating a discussion where each participant expresses their opinion about some statement. Here, both generative models produced thematically correct answers. However, the GroundHog model used more appropriate words, resulting in a response that was more consistent with the dialogue history. We hypothesize that this can be attributed primarily to the AMR input. For the first utterance, the AMR representation is as follows:

(miss :ARG0 (article) :ARG1 (trend :ARG1 (and) :ARG1-of (have-degree))) ... (process :ARG0 (artist) :ARG1 (trauma :poss a) :instrument (art :poss a))

Therefore, the main entities are "article", "trend", "artists", "trauma", and "art". The GroundHog model primarily relies on these words, whereas BART's response is primarily influenced by the word "generation". However, the frequent occurrence of "generation" does not capture the underlying meaning of the text.

General View We have determined that linguistic features individually demonstrate utility and yield interpretive results. There is also the potential for uncovering valuable hidden insights through their combination. Nevertheless, our research represents a step towards achieving a coherent and meaningful generation.

It is worth considering that linguistic features can also be manually specified when the current context is insufficient for parsers to accurately perform their tasks. Such manual specifications can facilitate dialogue management.

6 Conclusion and Future Work

In this paper, we investigated the efficacy of incorporating grounding and multi-grained linguistic information for multi-party conversation generation. To address the challenge of handling lengthy input texts, we proposed the GroundHog model, which leverages both grounding and linguistic features.

For evaluation, we collected a novel Reddit-based dataset designed for training dialogue systems. This dataset was augmented with linguistic features, including semantic and discourse information, as well as sentiment. Experiments involving both automatic metrics and human evaluation have shown that generated texts using linguistic inputs were more preferable. In our supplementary analysis, we interpreted the obtained results.

Further research directions include the investigation of other linguistic inputs as well as other representations of inputs. Also, we plan to experiment with the recent LLMs to analyze their possibilities of leveraging linguistic features.

Limitations

Our approach is not constrained by language and has the potential for universal application. At the same time, we introduce a novel Transformer architecture that ideally requires pre-training on a large dataset. Furthermore, the effectiveness of the methodology is constrained by the accuracy and reliability of the parsers used to extract linguistic features, as well as the performance of the grounding extraction model.

Ethics and Broader Impact

The use of large Transformer models for training has been linked to contributing to climate change. However, it is important to highlight that our research did not involve training these models from scratch. Instead, we conducted a fine-tuning process on pre-existing models.

As is the case with any generative model, it is not possible to ensure flawless quality in the generated output. At the same time, we do not make our model publicly available. We mitigate the risks associated with generation by filtering the dataset and making business logic modifications.

The presented dataset was collected from Reddit for the purpose of scientific research and subsequent analysis. It may exhibit certain inherent biases due to its specific origin, and we suggest using it for scientific purposes only.

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.

The publication was also supported by the grant for research centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730321P5Q0002) and the agreement with HSE University No. 70-2021-00139.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, LAW-ID@ACL 2013, August 8-9, 2013, Sofia, Bulgaria*, pages 178–186. The Association for Computer Linguistics.

Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1644–1650. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-amr: abstract meaning representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 684–695.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings* of Machine Learning Research, pages 2206–2240. PMLR.

Alexander Chernyavskiy and Dmitry Ilvovsky. 2023. Transformer-based multi-party conversation generation using dialogue discourse acts planning. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 519–529, Prague, Czechia. Association for Computational Linguistics.

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

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. *arXiv preprint arXiv:2011.06623*.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *CoRR*, abs/2209.14375.

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.

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.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialogue systems. *arXiv preprint arXiv:2005.12529*.

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.

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

Baber Khalid, Malihe Alikhani, Michael Fellner, Brian McMahan, and Matthew Stone. 2020. Discourse coherence, reference grounding and goal oriented dialogue. *arXiv preprint arXiv:2007.04428*.

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.

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

Yinquan Lu, Haonan Lu, Guirong Fu, and Qun Liu. 2021. KELM: knowledge enhanced pre-trained language representations with message passing on hierarchical relational graphs. *CoRR*, abs/2109.04223.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk*, 8:243 – 281.

OpenAI. 2023. Gpt-4 technical report.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. *arXiv preprint arXiv:2107.06963*.

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.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *AAAI*.

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.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Deeksha Varshney, Akshara Prabhakar, and Asif Ekbal. 2022. Commonsense and named entity aware knowledge grounded dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1335, Seattle, United States. Association for Computational Linguistics.

Junjie Wu and Hao Zhou. 2021. Augmenting topic aware knowledge-grounded conversations with dynamic built knowledge graphs. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 31–39.

Sixing Wu, Ying Li, Ping Xue, Dawei Zhang, and Zhonghai Wu. 2022. Section-aware commonsense knowledge-grounded dialogue generation with pretrained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 521–531.

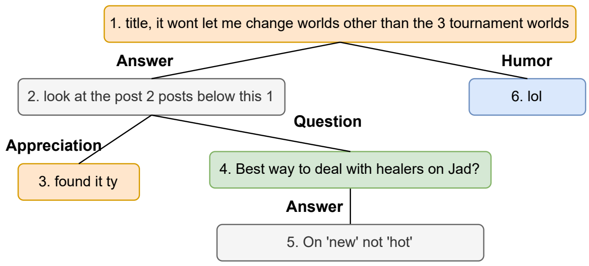

Figure 5: A manually annotated discourse tree for the multi-party dialogue. The color identifies the speaker, and edges indicate dialogue acts.

Bohao Yang, Chen Tang, and Chenghua Lin. 2023. Improving medical dialogue generation with abstract meaning representations. *arXiv preprint arXiv:2309.10608*.

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*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. *arXiv preprint arXiv:2010.08824*.

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.

A Linguistic Representations

Dialogue Acts Figure 5 illustrates an example of a multi-party conversation that has been annotated with dialogue acts. In this figure, each node in the graph represents an utterance in the conversation and includes attributes such as the speaker's identifier (represented by colors) and the utterance text. The edges connecting the nodes indicate the flow of conversation and include attributes such as the addressee, representing the recipient of the utterance, and the dialogue act, representing the specific function or purpose of the utterance.

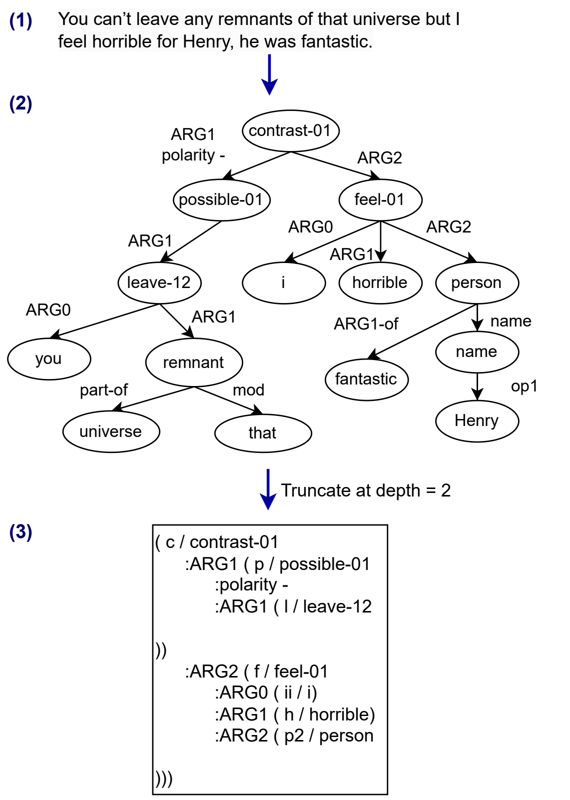

Figure 6: AMR representation for a single utterance and its truncated (by the first two levels) linearized representation. Here, (1) is the input text, (2) is the corresponding AMR graph, and (3) is the truncated plain graph2text representation.

Abstract Meaning Representation Figure 6 illustrates the representation of an utterance and its linearization using Abstract Meaning Representation (AMR). In this representation, words from the utterance are depicted as nodes in a graph, with edges representing the semantic relations between them. Higher-level vertices closer to the root of the graph capture the overall meaning, while lower-level vertices offer more specific details. In the given example, the core concept of contradiction is conveyed through the first two levels of the AMR graph. To enhance the efficiency of processing and reduce the length of the linearized representation, we only truncate the first levels of these graphs.