Imagination is All You Need!
Curved Contrastive Learning for Abstract Sequence Modeling Utilized on Long Short-Term Dialogue Planning

Anonymous ACL submission

Abstract

Motivated by the entailment property of multi-turn dialogues through contrastive learning sentence embeddings, we introduce a novel technique, **C**urved **C**ontrastive **L**earning (CCL), for generating semantically meaningful and conversational curved utterance embeddings that can be compared using cosine similarity. Inspired by the curvature of space-time (Einstein, 1921), we define the curved property of these embeddings as the semantic space curved by the relative turn distance (our time dimension) of utterance pairs. The resulting bi-encoder models can guide transformers as a response ranking model towards a goal in a zero-shot fashion by projecting the goal utterance and the corresponding reply candidates into a latent space. Here the cosine similarity indicates the distance/reachability of a candidate utterance toward the corresponding goal. Furthermore, we explore how these forward-entailing language representations can be utilized for assessing the likelihood of sequences by the entailment strength i.e. through the cosine similarity of its individual members (encoded separately) as an emergent property in the curved space. This allows us to imagine the likelihood of future patterns in dialogues, specifically by ordering/identifying future goal utterances that are multiple turns away, given a dialogue context. As part of our analysis, we investigate characteristics that make conversations (un)plannable and find strong evidence of planning capability over multiple turns (in 61.56% over 3 turns) in conversations from the DailyDialog (Li et al., 2017) dataset. Finally, we will show how we can exploit the curved property to rank one million utterance & context pairs, in terms of GPU computation time over 7 million times faster than DialogRPT (Gao et al., 2020), while being on average 2.8% qualitatively superior for sequences longer than 2 turns.

1 Introduction

Large Scale Transformers are becoming more and more popular in dialogue systems (Zhang et al. (2019), Peng et al. (2022)). Though these models are very effective in generating human-like responses in a given context, based on their learning objective to minimize perplexity, they tend to have trouble generating engaging dialogues (Gao et al., 2020). Meister et al. (2022) have shown that human conversations usually do not sample from the most likelihood of words like transformers do. We argue that one reason for this is that natural conversations can be (always) considered goal-oriented (even chitchat) and motivate this claim based on literature from psychology. These have shown that "Conversation is a goal-directed process" (Myllyniemi, 1986) as humans shift conversation topics based on the social connection/audience and use it to shape social relations (Dunbar et al., 1997). The psychological literature also elaborates on how humans are able to plan and simulate dialogues by utilizing inner speech as part of verbal working memory (Grandchamp et al., 2019).

> "Key to most of such models is that inner speech is posited as part of a speech production system involving predictive simulations or “forward models” of linguistic representations" (Alderson-Day and Fernyhough, 2015)

Keeping this in mind, we investigated dialogues under the aspect of "forward" entailing language representations by projecting them into a simple semantic sentence transformer (Reimers and Gurevych, 2019) latent space. We place a fixed position in the DailyDialog (Li et al., 2017) dataset as a goal utterance and measure the cosine similarity of the goal to every other utterance within the dialogue. Our own preliminary work revealed, as shown in figure 1, that the similarity of previous utterances to the goal utterance increases as they get closer to the goal utterance.

However, fluctuations between the speaker at the goal turn (saying the utterance later on) and

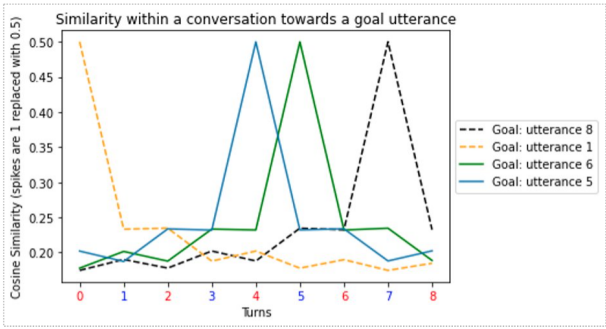

Figure 1: Entailment property of sentence transformer-based embeddings within conversations on DailyDialog

their dialogue partner can be observed. As we see on the blue & red highlighted turns, the goal turn speaker has a greater similarity to the goal utterance than the dialogue partner. We filtered all samples causing these fluctuations and find that these transitive entailing properties are essential for guiding the conversation toward the given goal. Regardless of whether the person had the intent to reach the target goal.We will demonstrate in this paper how we can build upon this phenomenon to generate semantically meaningful and conversational curved embeddings. In particular, by mixing the training objective of Natural Language Inference (NLI) for the semantic embedding space with a distance proportional and directional aware (through two special tokens [BEFORE] & [AFTER]) cosine similarity score of utterance pairs.

The resulting Curved Contrastive Learning (CCL) is presented on three tasks: **(1)** short-term planning, **(2)** next utterance selection, and **(3)** long-term planning.

(1) Short-term planning: CCL allows us to imagine the likelihood of a candidate utterance leading to a given goal utterance by projecting them together into one latent space (imaginary space). The cosine similarity indicates the distance/reachability of a candidate utterance towards the corresponding goal as illustrated in a transformer guidance example in figure 2. Thanks to the transitive property we can select the utterances at each turn greedily.

(2) Next utterance selection: The embeddings can be utilized for sequence modeling by only using the cosine similarity between the separately encoded sequence members. It is evaluated by the ranking performance of the human vs random utterances task given a dialog context.

(3) Long-term planning: Since these embeddings do not require entire sequences for sequence modeling, we can assess the likelihood of following patterns (of multiple goal utterances that are multiple turns apart) by using the entailment strength between these and the context in the curved space. We will evaluate this approach based on the ordering/identifying of future goal utterances.

Furthermore, we investigate two research questions:

- Do chit-chat conversations have planning capability? **(RQ1)**
- What characteristics make dialog planning possible? **(RQ2)**

The paper is structured as follows: In §2 we will discuss the related work. Following in §3 where we present the methodology, baselines as well as basic components for the advanced architectures. In §4 the short-term planning approaches, followed by the next utterance selection in §5 and the long-term planning approaches for ordering goals in §6. We will wrap up the paper with the experiments & discussion in §7 followed by the conclusion in §8.

2 Related Work

Our work builds upon two major concepts, dialogue planning, and entailment. Related publications from the stated fields are discussed below.

Dialogue Planning

While previously introduced planning techniques used several abstraction approaches (Teixeira and Dragoni, 2022), none of them exploited the characteristics of curved conversation embedding latent spaces. We argue that generating a complete dialogue path is unnecessary as we can simply choose the utterance in the transformer's search space that gets us closest to the goal at every turn. Ramakrishnan et al. (2022) proposed a similar idea on word level by applying constrained decoding to the dialogue response generation to increase the likelihood of a target word not only in the current utterance but also utterances in the future. Furthermore, DialogRPT (Gao et al., 2020) has been introduced as a dialogue response ranking model for depth, width, and upvotes prediction for utterance candidates. We will utilize DialogRPT as a baseline for our next utterance selection experiments based on the dialogue history.

Entailment

Entailment-based approaches have a long history in NLP and have been utilized for a lot of tasks as

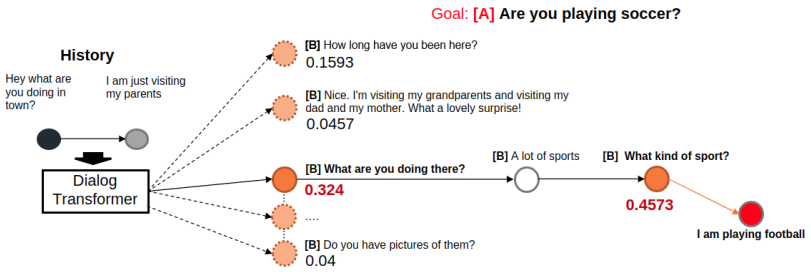

Figure 2: DialoGPT Guidance Example with Imaginary Embeddings with before `[B]` and after `[A]` token.

zero-shot classification tasks like relation extraction (Obamuyide and Vlachos, 2018) or zero-shot text classification (Yin et al., 2019). The idea of entailment graphs and making use of transitivity has been previously explored by Kotlerman et al. (2015) & (Chen et al., 2022). Textual entailment has also been applied to Dialogue Systems as an evaluation technique (Dziri et al., 2019) or for improving response quality through backward reasoning (Li et al., 2021). Contrastive learning with positional information has been previously applied to image segmentation (Zeng et al., 2021). While You et al. (2020) utilized contrastive learning with augmentations for graph neural networks (GNNs). Natural Language Inference (NLI) based transformers have been increasingly used for semantic textual similarity (STS) since the introduction of Sentence Transformers, thanks to bi-encoders (Reimers and Gurevych, 2019) that can compare sentence pairs with cosine similarity and therefore reduce computation time by a 234000 * fold. This trend has especially been supported by GPU Search (Johnson et al., 2017). These sentence transformers have successfully been applied to learn utterance representations for retrieving utterance replies in dialogue systems (Liu et al., 2021). However, without utilizing the curved property of conversations which we argue, as motivated in §1, is essential for forward representations.

3 Methods

In this section, we formally define the research questions (problem definition), our baselines for the evaluation, and the core of Imaginary Embeddings based on which advanced architectures will be built in the following sections.

*According to Reimers and Gurevych (2019) a set of 10000 Sentences would require 50 million inference computations with Bert which would, according to them, require 65 hours, while SBERT prior encoded would only take 5 seconds

3.1 Problem Definition Planning

As part of this paper, we will investigate two planning problems, short- and long-term planning. Short-term planning aims at guiding the conversation from the current position towards a given goal utterance g (which we define as a semantic utterance) over multiple turns. Long-term planning, on the other hand, targets the ordering/scheduling of a set of goals G (utterances that are multiple turns apart) within a conversation.

3.2 Long-Short Term Planning Evaluation

As part of this paper, we introduce a new evaluation technique, **L**ong-**S**hort **T**erm **P**lanning **E**valuation (**LSTPE**). LSTPE is split into Short- as well as Long-Term planning.

3.2.1 Short-Term Planing Evaluation

As part of the short-term planning evaluation, we evaluate the guidance capability of imaginary embeddings towards a given goal utterance. For this purpose, we split all dialogues within a given corpus $d \in C$ into subsets of $d[:h_l]$ which represents the history of utterances (or context) with a fixed length h_l, $d[h_l]$ the "correct" following utterance and $d[h_l + g_d]$ as goal utterance with a goal distance g_d. We then let a dialog transformer generate 100 candidate utterances given the context $d[:h_l]$ for every dialogue $d \in C$ which we project together with the goal utterance into the imaginary embedding. Following, we compare the ranking score of the original utterance to the artificially generated utterances.

3.2.2 Long-Term Planning Evaluation

Similar to the Short-Term planning, we take a corpus of dialogue data $d \in C$ and split it at fixed positions x into the dialogue history and three goal utterances $|G| = 3$. Given a dialogue history of length h_l, $\forall d \in C : d[:h_l], d[x], d[x + g_d], d[x + 2g_d]$ where $g_d \geq 2$ is the distance between the goals.

We define the first goal in distance as $x - h_l$ in the perspective of the dialogue history. The three resulting goal utterances result in 6 possible order permutations. Since 4 of them are partially ordered, we split the evaluation into ranking the partially ordered and reverse order to the true order separately. While this technique is simple and does not require any supervision, some samples due to the random selection will be without any context indistinguishable. E.g. an utterance like "oh, okay" could be at any position. Since all models are evaluated on the same data set, this is not an issue, however, an accuracy of 100% will be realistically not possible.

3.3 Next Utterance Selection Evaluation

Furthermore, we will test the embedding's capability of telling potential replies from random utterances given a dialog context by comparing it to DialogRPT (Gao et al., 2020) on a ranking task. The data set is built up in a similar way as for short-term planning.

3.4 Imaginary Embeddings with Curved Contrastive Learning

We introduce a novel self-supervised learning technique to generate semantically meaningful embeddings which have a conversation positional as well as directional awareness between utterance pairs. To generate these properties, we train a bi-encoder sentence transformer on two training objectives. The first objective builds upon the AllNLI dataset (a combination of SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2017)) with a simple Softmax Loss. To learn the graph structure of conversations, two special tokens `[BEFORE]` and `[AFTER]` are introduced. The model is (pre-)trained with a Cosine Similarity loss on DailyDialog (Li et al., 2017), by sliding through conversational data with a fixed length $l = 6$. Notably, we combine consecutive utterances of the same speaker. Based on this fixed length, the training data is constructed for a given window as follows:

$$\forall i \in \{1, .., l\} : \begin{cases} ([B]\ u[0], [A]\ u[i], s = \frac{l-i}{l} \\ ([B]\ u[i], [A]\ u[0],\ s = 0 \\ ([B]\ u[0], [A]\ u'[r], s = 0 \\ ([B]\ u'[r], [A]\ u[0], s = 0 \end{cases} \quad (1)$$

where $[A]$ = `[AFTER]`, $[B]$ = `[BEFORE]`, u the utterances in the observed window, u' a set of random utterances, and s the cosine similarity score. As 3.4 shows, the target cosine similarity for a positive sample pair is proportional to their positional distance in the dialogue (see illustration in figure 3). Three hard negatives are introduced, the first ensures the directional property by swapping the `[BEFORE]` and `[AFTER]` token. The following two are selected from a special dataset of random utterances. Figure 3 unveils the widespread utility of imaginary embeddings. As shown, we can simply pick the best candidate utterance for reaching a given goal by **imagining** the closeness of the candidate utterance to the goal in the curved space. Similar to an object in our universe that always moves on a straight line but is curved by space-time (Einstein, 1921), we can follow a line to our goal utterance by greedily selecting the best utterance on turn-to-turn bases. We illustrated this transitive property by the light red in-between nodes in figure 3. While in the short planning the candidate utterances are sampled from a dialog transformer, we can simply ignore the closeness of candidate utterances to the history. In long-term planning, however, we can exploit the curved property of context utterances for goal ordering as the next goal should be the closest to the context utterances. Analogous applies to the best next utterance in the next utterance selection/ranking task.

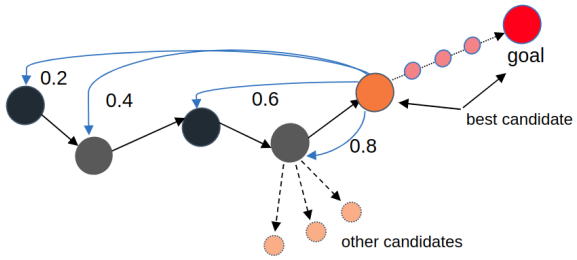

Figure 3: Curved property of Imaginary Embeddings. Grey/black nodes represent history utterances, orange nodes are utterance candidates, and dark orange is the best candidate as it is closest to the goal utterance (red). From the perspective of the best candidate encoded as `[A]`, the scores towards history illustrate the training objective as they are encoded with `[B]` tokens.

3.4.1 Adding Speaker Tokens

Furthermore, we can modify imaginary embeddings with additional speaker tokens. Given a multi-turn dialogue with two participants, the tokens `[O]` and `[E]` are added to the `[BEFORE]` utterance at the encoding step (for even and odd distances to the target utterance `[AFTER]`). Accord-

ingly, the training objective (see equation 3.4) for the curved property is slightly modified by adding hard negatives for false speaker matches.

4 Short Term Planning Approach (Transformer Guidance)

As described in section 3.2.1 we utilize imaginary embeddings as a re-ranking model. Respectively, we let a task-specific dialog transformer generate 100 candidate utterances given the context $d[:h_l]$ of a fixed length h_l for every sample dialogue $d \in C$. To get a diverse distribution of utterances we choose nucleus sampling with $p = 0.8$ and a temperature of $t = 0.8$. The generated utterances from the transformer are then projected in the imaginary embedding space and the goal similarity of $d[h_l + g_d]$ is measured. Following, we check the rank of the true utterance from the test set leading to the goal utterance. The average rank and the distribution of ranks within the dialogue are evaluated with respect to different history lengths h_l and different goal distances g_d.

5 Next Utterance Selection with Curving

Motivated by the curved property, the most suitable next utterance $u_f \in U_F$ for a dialogue sequence his should be closest to the individual utterances of the sequence on average. We can assess a relative likelihood between all future utterances by measuring the entailment strength P_E of every u_f to the history utterances based on the cosine similarity as follows:

$$P_E(u_f|his) = \sum_{u_i \in his} \frac{\mathbf{[B]\ u_i}\ \ \mathbf{[A]\ u_f}}{\|\mathbf{[B]\ u_i}\|\ \|\mathbf{[A]\ u_f}\|} \quad (2)$$

In the ranking evaluation, we will sort the results of $\forall u_f \in U_F : P_E(u_f|his)$ to determine the rank of the true utterance. Notably, we can observe the entailment strength (or activation) of individual utterances to a future one, which enables many other applications. Furthermore, we can utilize the curved context for greedily selecting the next goal $\max_{g \in G} P_E(g|his)$ in our long-term planning experiments. We will refer to this as greedy curving.

6 Long-Term Planning Approaches

In this section, we will describe how Imaginary Embeddings can be used to order goals (a set of utterances) within dialogues for long-term planning. The models are evaluated with **LSTPE**, a given set of goals G with $|G| = 3$, and an equal distance between each node.

6.1 Imaginary Embedding Chains

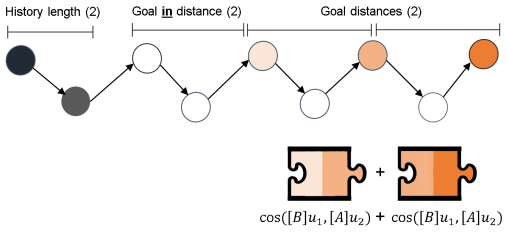

Figure 4: Long Term planning Dataset construction variables (history length, goal distances, (first) goal in distance) demonstrated. Furthermore, the concept of Imaginary Embedding Chains (IEC) is illustrated with its puzzle-like properties with the corresponding goal utterance colors.

Imaginary Embeddings are perfectly suited for this task as they can be concatenated into cosine similarity chains by using the (`[B]` before and `[A]` after token) as illustrated in figure 4. We mathematically define it as:

$$s(o) = \Big(\sum_{i \in o} \frac{\mathbf{[B]\ g_i}\ \ \mathbf{[A]\ g_{i+1}}}{\|\mathbf{[B]\ g_i}\|\ \|\mathbf{[A]\ g_{i+1}}\|} \Big) \quad (3)$$

where we choose the order of goals $o \in O$ by the highest similarity score s with $\max_{o \in O}(s(o))$ (strongest entailment strength) of a given sequence $o =< g_1, ..., g_n >$ of goals $g_i \in G$. While this chain can be arbitrarily long and, thanks to GPU tensor computations calculated rather quickly, the complexity with $\mathcal{O}(n!)$ for a brute force computation remains high.

6.2 Imaginary Embedding Chains with History Curving

Finally, we combine the concepts of Imaginary Embedding Chains and Curving by generating for every order $[g_1, g_2, g_3]$ a score (equation 4):

$$s(g_1, g_2, g_3) =< g_1, g_2, g_3 > + P_E(g_1|his) - \frac{1}{2} P_E(g_2|his) - P_E(g_3|his) \quad (4)$$

where $< g_1, g_2, g_3 >$ is the chain score of the given order and h is the history curving score for the corresponding goal. We motivate the addition of g_1 and the subtraction of g_3 (as well as g_2) based on the presumption that g_1 should be closest while g_3 should be the furthest away to the history with respect to the curved property.

7 Experiments

Our experiments are conducted on two dialog corpora, DailyDialog (Li et al., 2017) and the Microsoft Dialogue Challenge (MDC) corpus (Li et al., 2018). We experiment with two transformer architectures BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) to generate Imaginary Embeddings. In the short-term planning (transformer guidance) setting, we let our Imaginary Embeddings guide DialoGPT (Zhang et al., 2019) for DailyDialog and GODEL (Peng et al., 2022) for the MDC corpus. For the next utterance selection, we use DialogRPT (Gao et al., 2020) as a baseline.

7.1 Experimental Setup

While the DailyDialog data set has a test corpus of 1000 dialogues, we first have to generate a test data set for MDC. We do so by extracting the last 333 samples for each of the three task-oriented domains (movie-ticket booking, restaurant reservation, and taxi booking). This leaves us with 11,118 dialogues as training data for DailyDialog and 9088 training samples for MDC.

7.2 Self-Supervised Training

Apart from combining consecutive utterances of the same speaker and removing dialogues with utterances longer than 200 tokens, we apply no further pre-processing on the training data. As described in §3.4, we pre-train all our architectures in stage (1) with a mixed training objective of NLI and the Curved Contrastive Learning (CCL) on the DailyDialog corpus for 5 epochs. For all MDC models, we follow up with a second stage where we train on the target corpora with the curved property learning objective only for domain adaptation.While Long Term planning performs best after 5 epochs of further fine-tuning, short-term planning requires only between 0.5 to 1 epoch(s). We provide all model cards with a detailed description as part of our submission in an anonymous GitHub repository[†]. We will publish the models together with our training/evaluation scripts upon acceptance.

7.3 Evaluation Data sets

The evaluation data sets DailyDialog and MDC are constructed analogously. We construct the datasets for STP based on history length and goal in distance & LTP based on history length, goal **in** distance, goal distances respectively as illustrated in figure 4. Since MDC with an average number of 6.51 turns is even shorter than DialyDialog with 7.84, we are limited in the long-term planning to a shorter context as well as a goal in distance length.

7.4 Evaluation & Discussion

In the following sections, we will investigate how well these embeddings perform on our introduced **LSTPE** (§3.2) and on the next utterance selection task. In the main paper, we focus on our empirical findings and present the results of the experiments for space reasons in aggregated form. We provide a detailed analysis in the appendix, where we explore examples as well as demonstrate the curved property of dialogues in these embeddings. This is illustrated as vector chains in figure 7 or the average similarity of different distances and directions within dialogues (appendix A).

7.4.1 Short-Term Planning

As shown in the short-term planning aggregated results table 1, we split the results based on odd distance length (unveiling utterances of the dialog partner) and even distance (which would be uttered by the transformer). Both have at least 20% of the true candidate utterances in the top 5 (of 100) ranks, 50% in the top 25, and a max average rank of 32.56. We observe that speaker token-based imaginary embeddings on odd distances can even achieve 63% in the top 5 with the highest average rank of 14.01. This can be expected as odd utterances will be uttered by our dialog partner which we can greatly influence by our preceding utterance. Interestingly, we find that it is significantly easier to plan 3 turns ahead rather than 2 turns. This is portrayed in the detailed analysis based on the history length, goal distances, and the first goal distance (goal in distance) in table 3 (appendix). Our analysis unveils that the DailyDialog models have an advantage through their more diverse utterance distribution in selecting the true candidate utterance. Furthermore, they perform more consistently across different history lengths and goal distances. MDC, on the other hand, performs overall better but has a higher variance in its performance (with samples of different history lengths and goal distance). Concluding that the score distribution in the ranking process is either more strongly peaked (most in data sets with lots of request intents) or it more is flattened (especially on data with majorly

[†] https://anonymous.4open.science/r/ImaginationIsAllYouNeed-82BF

Goal in Distance	Human Utterance Ranking vs 100 utterances sampled from DialoGPT Large / GODEL Large (p=0.8, t=0.8)									
	Imaginary Embedding without Speaker Token					Imaginary Embedding with Speaker Token				
	Top 5 (in %)	Top 10 (in %)	Top 25 (in %)	Top 50 (in %)	Average Rank	Top 5 (in %)	Top 10 (in %)	Top 25 (in %)	Top 50 (in %)	Average Rank
DailyDialog Test Corpus										
Guidance **even** g distance	29.36	35.76	51.03	67.9	34.59	27.78	36.22	53.78	71.36	32.56
Guidance **odd** g distance	31.31	39.21	54.09	72.78	30.61	**63.49**	72.18	83.21	91.06	**12.9**
MDC Test Corpus										
Guidance **even** g distance	20.79	29.32	48.04	70.85	34.86	39.18	50.9	69.29	83.1	**22.09**
Guidance **odd** g distance	25.41	32.17	46.8	67.31	35.88	**63.06**	70.65	80.94	89.16	**14.01**

Table 1: Aggregated short-term planning evaluation for odd (unveiling utterances of the dialog partner) and even distances (which would be uttered by the transformer itself).

inform intents). We explore this in detail in the appendix C. This flattened score distribution can be expected as in many cases of providing information, the actual information has little impact on future turns considering a structured task-oriented setting (e.g. replying on how many people will attend a reservation).

7.4.2 Next Utterance Selection based on Curved History

The sequence modeling capability is evaluated based on the normalized average rank (of the true following utterance compared to all other utterances at the same position of the corresponding corpus). We find that the DailyDialog corpus clearly outperforms MDC across all variations. As we demonstrate in figure 5, DailyDialog performs best with an average rank in the top 10% over all history lengths (the entire history projected in the curved space with speaker tokens). For sequences longer than 2 turns, it even outperforms all our base variants of DialogRPT (human vs. random) by at least 2.8%. Overall, we find that DialogRPT has trouble with increasing sequence lengths as input and find that keeping the last two utterances performs best. Notably, we can assess the entailment strength of 1000000 dialogue paths (1000 dialogues $\times$1000 utterances) as described in equation 4 in terms of GPU computation time over 7 million times faster than DialoRPT which we explore in more detail in the appendix B.1. While our experiments on MDC for the next utterance selection show weak results, in summary, MDC shows the same fluctuations between primarily inform & requests intents. While the ranking approaches based on only the last utterance are most of the time superior, we observe on odd turns (where we have a lot of request intents) the entire history usually performs better relative to even distances. Conversely, we notice that approaches based on only the last utterance are especially good on turns where we see more informing intents (replying to the request). We further explore this in the appendix B.2.

7.4.3 Long Term Planning Evaluation

The short turn length of the two corpora becomes especially troublesome in the long-term planning evaluation. Here, we are limited to short context/history lengths as well as short goal distances and (first) goal **in** distances.Across all models and datasets, we observe a solid average rank of 1.67 (between 1 and 2 for all approaches) on identifying the correct order of 3 goal utterances within their 6 possible orders as table 2 unveils. While our MDC embeddings had especially trouble with utterance selection in width (selecting an utterance from the same dialog depth §7.4.2), we find that MDC shows a stronger performance on greedy goal selection (Greedy Curving (GC)) on classic embeddings thanks to the solidified sequential structure of task-oriented dialogues. This advantage lets MDC outperform DailyDialog also on all other approaches. When Speaker tokens come into play, however, MDC drops while DailyDialog improves in performance compared to classic imaginary embeddings. Imaginary Embedding Chains (IEC) and with curved context (IEC & CU) show similar performance in aggregated form. However, when the context is close (i.e. the first goal is not far away) IECs with a curved context prevail. This changes with increasing distance of goals or first goal **in** distance as highlighted in table 4 of the appendix. Here, IECs with no context keep an advantage. In terms of the MDC planning capability, the perfor-

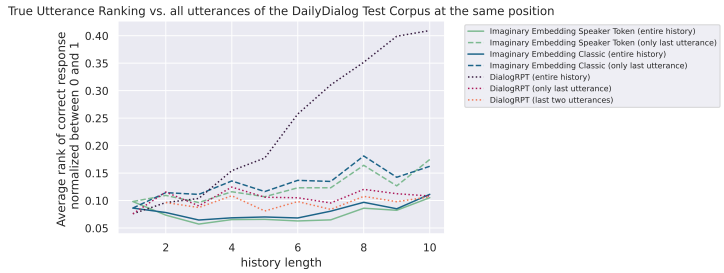

Figure 5: Normalized average rank of next utterance selection based on dialog history on DailyDialog. Demonstrated are different Curving variants (only the last utterance or the entire history), classic as well as Speaker Token-based embeddings, and the DialogRPT pre-trained on the human vs random utterance task baseline.

	Imaginary Embedding w.o. Speaker Token						**Imaginary Embedding with Speaker Token**					
	partially ordered				**Reverse order**		**partially ordered**				**Reverse order**	
Model	**Top 1 (in %)**	**Top 2 (in %)**	**Top 3 (in %)**	**Top 4 (in %)**	**Top 1 (in %)**	**Average Rank**	**Top 1 (in %)**	**Top 2 (in %)**	**Top 3 (in %)**	**Top 4 (in %)**	**Top 1 (in %)**	**Average Rank**
DailyDialog Test Corpus												
IEC	49.99	70.62	85.26	93.42	79.17	1.8	51.60	72.22	86.82	94.94	81.18	1.78
IEC & CU	50.69	71.24	85.09	93.63	78.54	1.79	51.07	72.98	86.9	94.97	79.87	1.78
GC	57.87	82.47	-	-	-	1.6	57.32	83.89	-	-	-	1.59
MDC Test Corpus												
IEC	58.72	77.43	90.28	96.38	85.28	1.65	56.83	77.50	90.19	95.44	84.52	1.65
IEC & CU	61.59	77.72	90.15	96.79	86.25	1.63	58.63	78.62	91.20	95.72	85.44	1.62
GC	66.30	89.61	-	-	-	1.44	56.05	80.59	-	-	-	1.64

Table 2: Aggregated Long-Term Planning Evaluation on 3 goals with ((2, 2, 2), (2, 2, 0) and (2, 2, 1)) with (history length, goal distances, first goal **in** distance). Models include Imaginary Embedding Chain (IEC), Imaginary Embedding Chain + Curving (IEC & CU), and Greedy Curving (GC).

mance drop-off between the two most common intents, request and inform, is similar, although not as severe as in short-term planning or the next utterance selection.

8 Conclusion

In this paper, we introduced Curved Contrastive Learning, a novel technique for generating forward-entailing language embeddings. We demonstrated that these can be utilized on various sequence modeling tasks by only using the cosine similarity between the separately encoded sequence members in the curved space. In particular, for the next utterance selection based on the curved history of utterances (where DailyDialog's true utterances are constantly in the top 10%), outperforming DialogRPT on sequences longer than 2 turns while in terms of GPU computation being over 7 million times faster. Furthermore, we have shown their pattern recognition ability on the ordering/identification of future representations even at longer distances and far apart (with an average rank of 1.67/6). We also demonstrated that these embeddings can be applied to guiding dialog transformers to approach a goal over multiple turns. In particular, by imagining the closeness of candidate utterances towards the goal through the transitive properties of the curved space. Following up on our claim, that even chit-chat can be considered goal-oriented **(RQ1)**, we find strong evidence of planning capability in chit-chat conversations over multiple turns. E.g. 48.83% / 61.56% (within the top 5 / top 10 utterances in the re-ranking) on 3 turns ahead. Our **RQ2** can be answered by the fact that we observe significant differences in the plannability of different intents. Our empirical analysis shows that request intents are significantly easier to plan than informing intents. While our focus in this paper was mainly on the introduction of Imaginary Embeddings and their utilization to dialogue planning, we leave much more space for further evaluation, analysis, and applications on the curved properties of our ~~universe~~ [‡] embeddings in future works.

[‡]In tribute to our fellow researchers in the field of physics for their inspiring work on the curvature of spacetime

9 Limitations

One of our limitations is that the data is split for short-term planning and long-term planning at fixed positions which on one side shows the overall planning capability on different datasets unbiasedly but on the other hand mixes the planning ability of the datasets with the overall performance of the embeddings. We have demonstrated in section C.2 that this can lead in many cases to unplannable examples. While this means that our embeddings should overall perform better than our results suggest, in the future, we should create either a human-filtered dataset where planning is always possible or either create a human benchmark as a further baseline. Furthermore, we rely in short-term planning (transformer guidance) on the generated utterance distributions by transformers where we have to balance between semantic diversity and the likelihood of utterances. We control these with temperature and nucleus sampling (top p) and found the best trade-off with a temperature of 0.8 and a top p of 0.8. Nonetheless, this can still lead to utterances that might lead to the goal but that would be not considered by humans as very likely based on the given context as we explore in C.2. Furthermore, in the next utterance selection, we utilize the vanilla DialogRPT which has been evaluated in the original paper (Gao et al., 2020) on DailyDialog but seemingly was not trained on a task-oriented corpus. Since we find that the next utterance selection based on the curved property of the context in a task-oriented setting like MDC is almost always worse than just taking the last utterance, we have not taken any further steps for experiments on our baseline DialogRPT in this domain.

10 Ethics

Like other language models, our model is prone to bias from training data sets (Schramowski et al., 2022)(Mehrabi et al., 2019). This is something to keep in mind when fine-tuning the model for domain adaptation. Since the models are for guidance only, we do not see any direct threats related to language generation. Still, if an individual intentionally wants to harm others and trains a language model to generate harmful utterances, our model could be employed to support this process. In contrast, however, we argue that these embeddings have great potential through their transitive properties to foresee and deflect harmful utterances from afar. Considering the risk that language models pose to humans (Weidinger et al., 2021), these embeddings could be utilized as a filter on top of generative language models, e.g. removing utterances that would increase the probability of leading to an utterance of a large set of harmful utterances. Our proposed model has a relatively small model size and shows higher efficiency during training & inference compared to DialogRPT, therefore we see great potential for reducing the carbon footprint in utterance retrieval tasks, in accordance with recent efforts in NLP (Strubell et al., 2019) (Patterson et al., 2021).

References

Ben Alderson-Day and Charles Fernyhough. 2015. Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, 141:931 – 965.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Zhibin Chen, Yansong Feng, and Dongyan Zhao. 2022. Entailment graph learning with textual entailment and soft transitivity.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Robin Dunbar, Anna Marriott, and Neill Duncan. 1997. Human conversational behavior. *Human nature (Hawthorne, N.Y.)*, 8:231–246.

Nouha Dziri, Ehsan Kamalloo, Kory W. Mathewson, and Osmar R. Zaïane. 2019. Evaluating coherence in dialogue systems using entailment. *CoRR*, abs/1904.03371.

Albert Einstein. 1921. *Relativity: The Special and General Theory*. Routledge.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. *CoRR*, abs/2009.06978.

Romain Grandchamp, Lucile Rapin, Marcela Perrone-Bertolotti, Cédric Pichat, Célise Haldin, Emilie Cousin, Jean-Philippe Lachaux, Marion Dohen, Pascal Perrier, Maëva Garnier, Monica Baciu, and Hélène Loevenbruck. 2019. The ConDialInt Model: Condensation, Dialogality, and Intentionality Dimensions of Inner Speech Within a Hierarchical Predictive Control Framework. *Frontiers in Psychology*, 10:2019.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *CoRR*, abs/1702.08734.

Lili Kotlerman, Ido Dagan, Bernardo Magnini, and Luisa Bentivogli. 2015. Textual entailment graphs. *Natural Language Engineering*, 21:699 – 724.

Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2021. Improving response quality with backward reasoning in open-domain dialogue systems. *CoRR*, abs/2105.00079.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. Dialoguecse: Dialogue-based contrastive learning of sentence embeddings.

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.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical decoding for natural language generation. *CoRR*, abs/2202.00666.

Rauni Myllyniemi. 1986. Conversation as a system of social interaction. *Language & Communication*, 6(3):147–169.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *CoRR*, abs/2104.10350.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. arXiv.

Ramya Ramakrishnan, Hashan Buddhika Narangodage, Mauro Schilman, Kilian Q. Weinberger, and Ryan McDonald. 2022. Long-term control for dialogue generation: Methods and evaluation.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *CoRR*, abs/1906.02243.

Milene Teixeira and Mauro Dragoni. 2022. A review of plan-based approaches for dialogue management. *Cognitive Computation*, 14.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *CoRR*, abs/1909.00161.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pages 5812–5823. Curran Associates, Inc.

Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, Haiyun Yuan, Meiping Huang, Jian Zhuang, Jingtong Hu, and Yiyu Shi. 2021. Positional contrastive learning for volumetric medical image segmentation. *CoRR*, abs/2106.09157.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation.

A Imaginary Embedding extended analysis

We analyze the Imaginary Embeddings based on their average similarity to different distances of utterances pairs within dialogues as well as their direction as shown in figure 6. While the model's average similarity is far from the training objective, the scores show a favorable decay considering the

distance for positive examples as well as a relatively low similarity for false direction utterance pairs. Furthermore, we have illustrated the curved

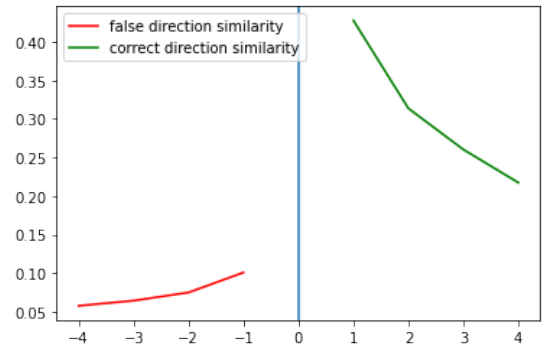

Figure 6: Average Imaginary Embedding Similarity to correct and false direction utterances based on turn distance on DailyDialog Test Corpus

property of these embeddings as directed graphs of dialogues in figure 7 where we notice a tendency of utterances at the beginning of the dialogue in the close right and the last utterance (encoded with the after token) deeper on the left.

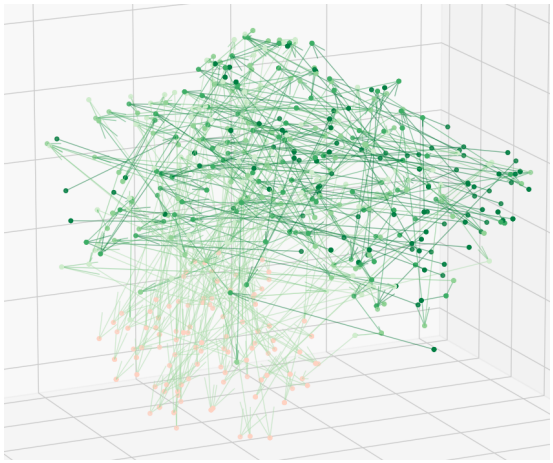

Figure 7: t-SNE visualization of first 4 utterances of the first 100 dialogues of the DailyDialog Test Corpus in curved Embedding Space. From Dark green to light green ($u_1 \rightarrow u_2 \rightarrow u_3$) nodes as well as edges encoded with the `[BEFORE]` token to u_4 encoded with `[AFTER]` token as light red.

B Next Utterance Selection Extended Analysis

For the next utterance selection we provide an extended description for our speed comparison as well as the MDC results.

B.1 Speed Comparison

In this section, we will investigate the speed of our introduced curving technique to rank 1000 incoming utterances for 1000 different contexts by comparing it to DialogRPT. Thanks to the curved property we have to only encode the new incoming utterances with the after token which takes on GPU (`A100-40GB`) around one second (1.16s). Following, we load the tensors on GPU, in particular, the history tensor H (encoded with the after tokens) with $(batch, h\_len, emb)$ (here with a batch size of 1000 and a history length of 3) as well as the following utterance candidates U(with $(batch, n\_cand, emb)$ which takes 0.091 seconds. Since we got normalized embeddings from the sentence transformer we can compute the cosine similarity-based score for the 1000000 dialogues in **one** simple batch matrix multiplication $U \odot H.T$ by transposing the history with dimensions $(1, 2)$. Following we sum across the second dimension (history dim) like equation 4 illustrates. Compared to DialogRPT on the same `A100-40GB` GPU, these two tensor computations (batch matrix multiplication & sum through the history dimension) take only 0.000531 seconds for a history length of 3 while DialogRPT needs for the same task, with a batch size of 32, 4023.78 seconds (67 minutes). As $\frac{4023.78}{0.000531} = 7.58 \cdot 10^6$ we conclude a 7 million time faster GPU computation time. Thereafter, we sort the similarity scores and search the index of the true utterance and return the corresponding rank which takes around 0.212 seconds.

B.2 MDC Results

We demonstrate the results of the MDC next utterance selection in figure 8 where we observe as described in the main paper the symmetry between inform and request intents either profiting from only the last utterance or the entire history.

C Extended Short-Term Planning Evaluation

As part of the extended Short Term Planning Evaluation, we investigate the extended results based on the history length, goal distances, and the first goal distance (goal in distance) in table 3 and demonstrate examples.

C.1 Detailed Short-Term Planning Evaluation

Table 3 unveils that additional speaker tokens show improvement in the MDC Test corpus across all

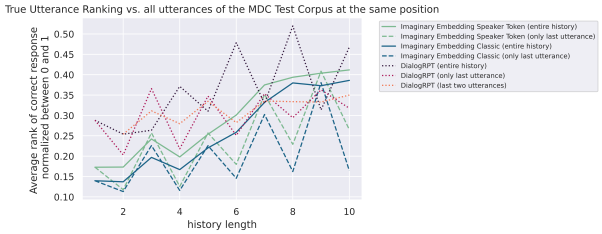

Figure 8: Normalized average rank of next utterance selection based on dialog history on MDC. Demonstrated are different Curving variants (only the last utterance or the entire history), classic as well as Speaker Token-based embeddings, and the DialogRPT pre-trained on the human vs random utterance task baseline.

tested categories. While classic embeddings show on MDC a similar performance across all even distances, we can observe two spikes at position $(3, 1)$ and $(5, 1)$ with (h_l, g_d) on odd distances with 51.17% / 45.80% in the top 5 respectively. At these positions, we monitor a 33% increase in the standard deviation on average of the distribution of guidance scores i.e. that the model is much more decisive in its ranking. We analyzed the intent at these positions and find a two times increase in requests and a 38% decrease in inform intents to the data set's average. While the speaker token-based embeddings show that we can overcome this gap for odd distances, we still find that the two lowest performers on $(4, 1)$ & $(4, 3)$ with "only" 53.03% & 51.45% in the top 5 have all a minimum of 80% of informing intents. Since the two corpora use separate latent spaces, we do not compare them on a simple standard deviation. Instead, we take the sum of average standard deviations as a baseline and divide it by the sum of the standard deviations (for each data set) of the standard deviations (for each transformer utterance distribution) to measure the variation in performance over different testing parameters history length, goal distances, (first) goal in distance. With a 35% higher score, DailyDialog shows less variance through different test parameters. Nonetheless, we find that DailyDialog has a 12% higher semantic variance across all utterances in the transformer-generated distributions than MDC by measuring their average semantic similarity with a simple semantic sentence transformer.

C.2 Examples of Short-Term Planning

While we provide construction of our evaluation datasets, we still want to highlight some of the strengths and weaknesses of our introduced embeddings. In the example on the left of figure 9, we can see that without knowing what the person is going to say, the model can sometimes move toward the goal too greedily. In the example on the right, we see that the model can also understand more complex relations, where the only way to get to a conversation state where someone would utter "look behind you. They are coming this way" would be in a manner of playing catch me as the model ranks it on the first position. A lot of the weaker ranking results are due to the fixed split of data as demonstrated in figure 10. We observe in the first example (left) that the model tries to unveil the utterance "You're right" by trying to get the other person into an argument (rank 1) where it hopes the person would then agree to their own opinion 3 turns later or by trying to unveil the utterance right away (rank 2). In the example in the middle, we see the drawback of purely relying on the transformer's context-aware utterance generation as the selected utterance of "pint of wine" might be closer to fruits than beer but at the same time is not a valid answer. This can be also observed in the last example (right).

Embedding Type	History Length	Goal Distance	n	Human Utterance Ranking vs 100 utterances sampled from DialoGPT Large / GODEL Large (p=0.8, t=0.8)									
				Imaginary Embedding without Speaker Token					Imaginary Embedding with Speaker Token				
				Top 5 (in %)	Top 10 (in %)	Top 25 (in %)	Top 50 (in %)	Average Rank	Top 5 (in %)	Top 10 (in %)	Top 25 (in %)	Top 50 (in %)	Average Rank
DailyDialog Test Corpus													
Guidance with **even** goal distance g_d (saying goal by yourself)	2	2	741	23.08	31.44	50.74	70.31	33.65	24.70	33.87	55.06	75.57	30.28
	2	4	534	23.03	31.65	48.13	66.85	35.61	22.10	32.02	51.31	71.72	32.57
	5	2	479	25.05	31.52	44.47	63.47	38.03	20.88	29.23	49.69	69.52	34.87
	5	4	323	15.79	22.60	39.01	56.66	43.02	17.65	24.15	42.11	66.25	38.27
	10	2	102	48.04	51.96	60.78	77.45	27.18	36.27	45.10	61.76	70.59	30.37
Guidance with **odd** goal distance g_d (unveiling goal utterance in dialogue partner)	2	1	918	42.37	50.54	66.88	84.64	21.74	70.59	78.54	87.15	94.55	9.15
	2	3	651	23.66	33.33	51.00	71.89	32.05	52.53	60.52	74.04	84.79	19.19
	5	1	534	35.02	43.26	58.61	76.40	27.90	67.79	77.53	86.70	93.26	10.46
	5	3	385	18.44	23.64	40.00	61.04	40.92	48.83	61.56	76.62	85.97	18.83
	10	1	183	36.61	44.81	54.10	69.95	30.49	77.60	82.51	91.26	96.72	6.86
MDC Test Corpus													
Guidance with **even** goal distance g_d (saying goal by yourself)	2	2	600	20.67	28.83	43.00	64.33	37.68	45.83	55.00	69.33	84.33	20.41
	2	4	417	21.58	31.18	47.00	67.63	36.02	47.48	55.16	70.26	83.45	20.85
	3	2	545	22.02	32.66	50.64	69.72	33.33	34.68	44.40	66.24	78.35	25.08
	3	4	344	26.16	38.08	53.49	77.62	28.97	41.28	53.20	67.44	85.76	20.93
	4	2	417	20.62	29.50	46.28	64.99	36.58	37.89	47.96	67.63	85.13	21.06
	4	4	234	16.67	23.08	47.01	70.51	37.24	40.60	53.42	73.93	89.74	18.04
	5	2	344	18.02	24.42	40.70	60.47	40.94	29.36	41.86	61.05	77.03	26.79
	5	4	161	20.50	34.78	56.52	78.26	28.09	44.72	58.39	75.78	88.82	17.32
Guidance with **odd** goal distance g_d (unveiling goal utterance in dialogue partner)	2	1	893	20.83	27.32	40.54	61.59	38.89	63.83	69.99	81.41	90.26	13.46
	2	3	545	31.19	38.53	55.41	73.76	29.92	69.91	77.06	83.30	90.28	11.78
	3	1	600	**51.17**	58.00	70.33	82.00	20.75	69.17	74.17	83.33	91.50	12.03
	3	3	417	15.83	25.18	43.88	68.35	37.87	67.39	73.62	83.93	93.29	11.25
	4	1	545	18.17	26.06	43.30	67.16	36.16	53.03	63.49	76.70	84.04	18.23
	4	3	344	17.44	25.58	42.44	61.34	39.51	51.45	62.50	76.16	83.14	18.42
	5	1	417	**45.80**	52.28	63.07	74.34	26.56	73.38	77.22	85.85	91.85	10.85
	5	3	234	16.24	19.23	32.91	58.55	46.47	71.37	77.78	88.46	92.74	9.92

Table 3: Detailed Short-Term Planning Evaluation with n (number of evaluation samples)

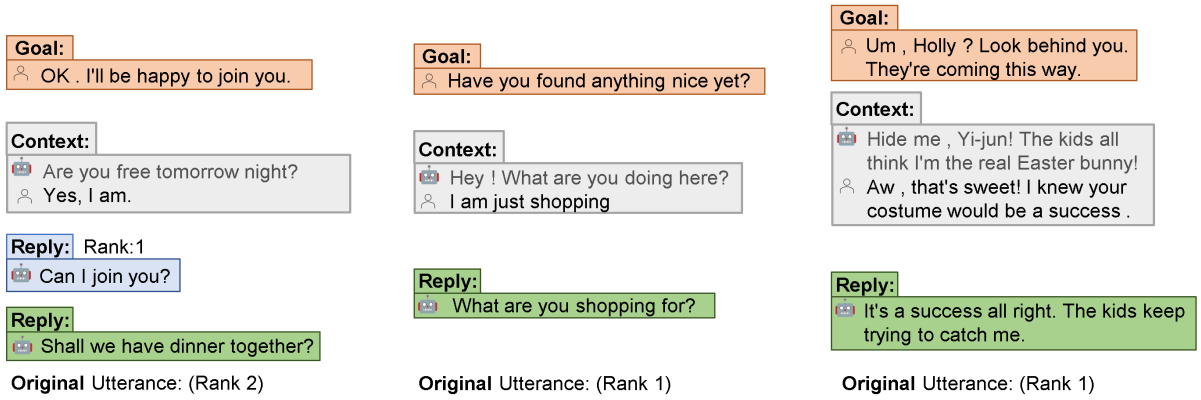

Figure 9: Good Ranking Examples on DailyDialog Test Corpus with a history length of 2 and a goal distance of 3. The goal in red, the context in grey, the true utterance in green, and the transformer-generated utterance in blue

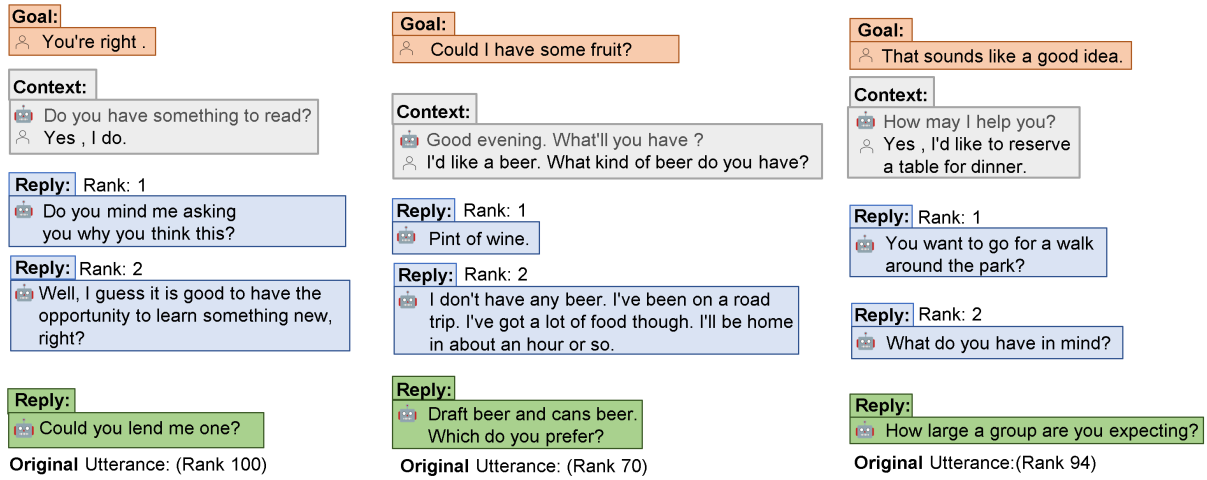

Figure 10: Bad Ranking Examples on DailyDialog Test Corpus with a history length of 2 and the goal distance of 3. The goal in red, the context in grey, the true utterance in green, and the transformer-generated utterance in blue

D Long-Term Planning Results

We present our detailed Long Term planning results in table 4 as well as examples in the following subsection.

D.1 Long-Term Planning Examples

Alike for short-term planning, we will demonstrate examples to present the weaknesses and as well as strengths of the embeddings. In figure 11 we show two very easy examples, where we can follow the conversation well without knowing the replies of the other dialogue partner. This changes especially in figure 12 where in the left example it is also for us very difficult to order the corresponding utterances. While one could argue that emergency calls tend to start with the location of the incident, the utterance "I haven't checked yet" makes the ordering of the utterances without any further context very difficult. This can also be observed in the right example of figure 12, however, one could argue that based on the context to which both IEC+CU and GC have access, the predicted order (of these two) makes more sense than the original reply order. Nonetheless, both examples show that some of these orders are debatable.

					LTP Planning Evaluation for 3 Goals											
					Imaginary Embedding without Speaker Token						Imaginary Embedding with Speaker Token					
					partially ordered				Reverse order		partially ordered				Reverse order	
Model	**History Length**	**Goal Distances**	**First Goal In Distance**	n	**Top 1 (in %)**	**Top 2 (in %)**	**Top 3 (in %)**	**Top 4 (in %)**	**Top 1 (in %)**	**Average Rank**	**Top 1 (in %)**	**Top 2 (in %)**	**Top 3 (in %)**	**Top 4 (in %)**	**Top 1 (in %)**	**Average Rank**
DailyDialog Test Corpus																
	2	2	0	385	57.66	72.47	87.79	93.25	81.56	1.67	58.70	76.10	91.43	97.14	84.16	1.66
	2	2	1	323	46.13	68.73	84.83	94.74	79.26	1.89	51.39	70.90	86.07	94.43	81.42	1.79
	2	2	2	230	46.52	67.83	83.04	90.87	74.78	1.83	46.96	67.39	83.91	92.17	78.70	1.85
IEC	2	2	3	183	**44.26**	66.12	77.60	91.26	73.22	1.94	50.82	68.85	83.61	93.99	77.60	1.84
	4	2	0	230	46.52	67.83	83.04	90.87	74.78	1.83	46.96	67.39	83.91	92.17	78.70	1.85
	4	2	1	183	44.26	66.12	77.60	91.26	73.22	1.94	50.82	68.85	83.61	93.99	77.60	1.84
	4	2	2	102	**43.14**	68.63	82.35	92.16	73.53	1.89	39.22	60.78	79.41	93.14	64.71	2.07
	2	4	0	51	37.25	56.86	78.43	86.27	76.47	2.00	47.06	66.67	96.08	98.04	84.31	1.86
	2	2	0	385	57.92	76.36	90.13	95.84	84.42	1.66	**59.74**	77.92	92.21	98.18	85.45	1.66
	2	2	1	323	46.75	67.80	85.14	95.05	77.71	1.90	**47.37**	70.59	85.45	94.12	78.95	1.84
	2	2	2	230	**47.39**	69.57	80.00	90.00	73.48	1.81	46.09	70.43	83.04	92.61	75.22	1.85
IEC & CU	2	2	3	183	44.26	63.39	77.60	92.35	66.12	1.99	45.90	62.84	79.78	93.44	71.04	1.98
	4	2	0	230	50.87	69.57	82.61	94.35	73.48	1.85	**50.87**	75.65	86.52	95.22	74.78	1.76
	4	2	1	183	**47.54**	68.31	83.06	94.54	71.04	1.90	53.55	72.68	81.42	92.90	69.95	1.76
	4	2	2	102	42.16	66.67	83.33	94.12	71.57	1.96	40.20	59.80	78.43	93.14	66.67	2.08
	2	4	0	51	41.18	74.51	84.31	92.16	78.43	1.83	**56.86**	86.27	90.20	96.08	84.31	1.57
	2	2	0	385	70.13	88.05	-	-	-	1.42	70.13	88.83	-	-	-	1.41
	2	2	1	323	56.97	83.28	-	-	-	1.60	51.39	81.11	-	-	-	1.67
	2	2	2	230	46.52	76.09	-	-	-	1.77	50.43	81.74	-	-	-	1.68
GC	2	2	3	183	48.09	74.32	-	-	-	1.78	44.81	73.77	-	-	-	1.81
	4	2	0	230	62.61	86.52	-	-	-	1.51	63.48	85.65	-	-	-	1.51
	4	2	1	183	50.82	82.51	-	-	-	1.67	56.28	84.15	-	-	-	1.60
	4	2	2	102	45.10	74.51	-	-	-	1.80	39.22	75.49	-	-	-	1.85
	2	4	1	51	78.43	90.20	-	-	-	1.31	82.35	90.20	-	-	-	1.27
MDC Test Corpus																
	2	2	0	234	52.99	79.91	90.60	97.01	85.47	1.70	50.85	74.79	89.74	96.15	83.33	1.76
IEC	2	2	1	161	66.46	78.88	91.93	95.65	86.34	1.52	67.08	82.61	91.93	95.03	88.20	1.46
	2	2	2	106	48.11	72.64	88.68	95.28	81.13	1.80	47.17	71.70	85.85	94.34	79.25	1.83
	3	2	0	161	66.46	78.88	91.93	95.65	86.34	1.52	67.08	82.61	91.93	95.03	88.20	1.46
	3	2	1	106	48.11	72.64	88.68	95.28	81.13	1.80	47.17	71.70	85.85	94.34	79.25	1.83
	3	2	2	75	**56.00**	81.33	92.00	96.00	82.67	1.61	56.00	81.33	92.00	94.67	88.00	1.58
	2	2	0	234	**65.81**	86.32	93.59	97.86	93.16	1.49	60.68	82.48	94.87	98.72	92.31	1.59
IEC & CU	2	2	1	161	**67.08**	77.02	90.06	96.27	84.47	1.57	65.22	80.75	90.06	95.03	85.71	1.52
	2	2	2	106	**51.89**	69.81	86.79	96.23	81.13	1.83	50.00	72.64	88.68	93.40	78.30	1.74
	3	2	0	161	**68.32**	80.12	93.17	96.27	85.71	1.49	52.80	75.78	82.61	95.65	80.75	1.79
	3	2	1	106	**50.94**	68.87	85.85	95.28	80.19	1.84	42.45	61.32	77.36	91.51	81.13	1.02
	3	2	2	75	46.67	66.67	81.33	94.67	78.67	1.94	28.00	50.67	73.33	85.33	58.67	2.22
	2	2	0	234	81.20	95.73	-	-	-	1.23	76.92	95.30	-	-	-	1.28
GC	2	2	1	161	67.70	88.20	-	-	-	1.44	45.96	79.50	-	-	-	1.75
	2	2	2	106	50.00	84.91	-	-	-	1.65	45.28	66.98	-	-	-	1.88
	3	2	0	161	72.67	90.06	-	-	-	1.37	39.13	69.57	-	-	-	1.91
	3	2	1	106	46.23	83.96	-	-	-	1.70	48.11	67.92	-	-	-	1.84
	3	2	2	75	45.33	72.00	-	-	-	1.83	24.00	41.33	-	-	-	2.35

Table 4: Detailed Long-Term Planning Evaluation with n = number of evaluation samples

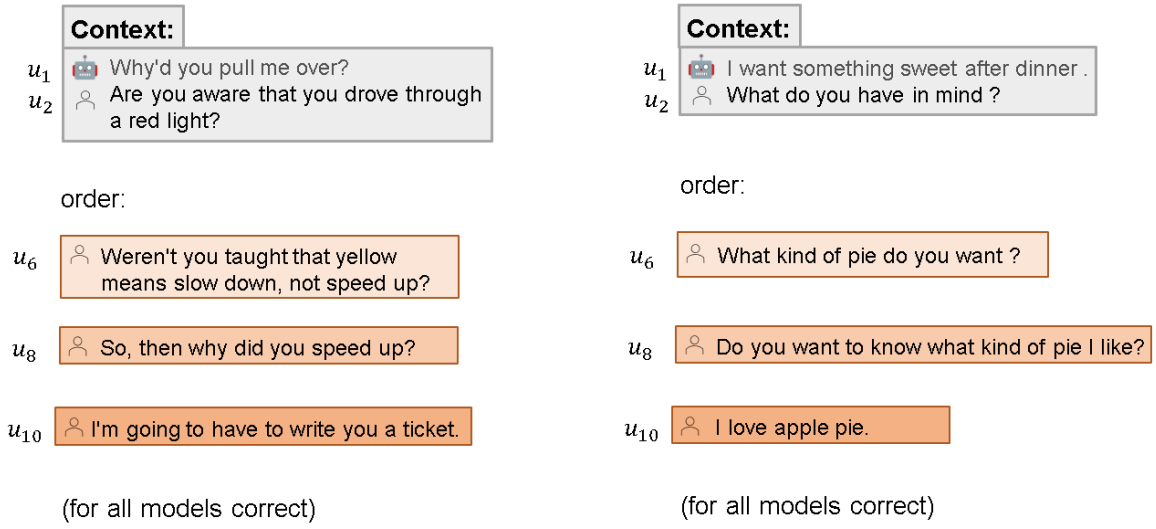

Figure 11: Bad Ranking Examples on DailyDialog Test Corpus with history length of 2, the goal distance of 2, and goal in distance of 3

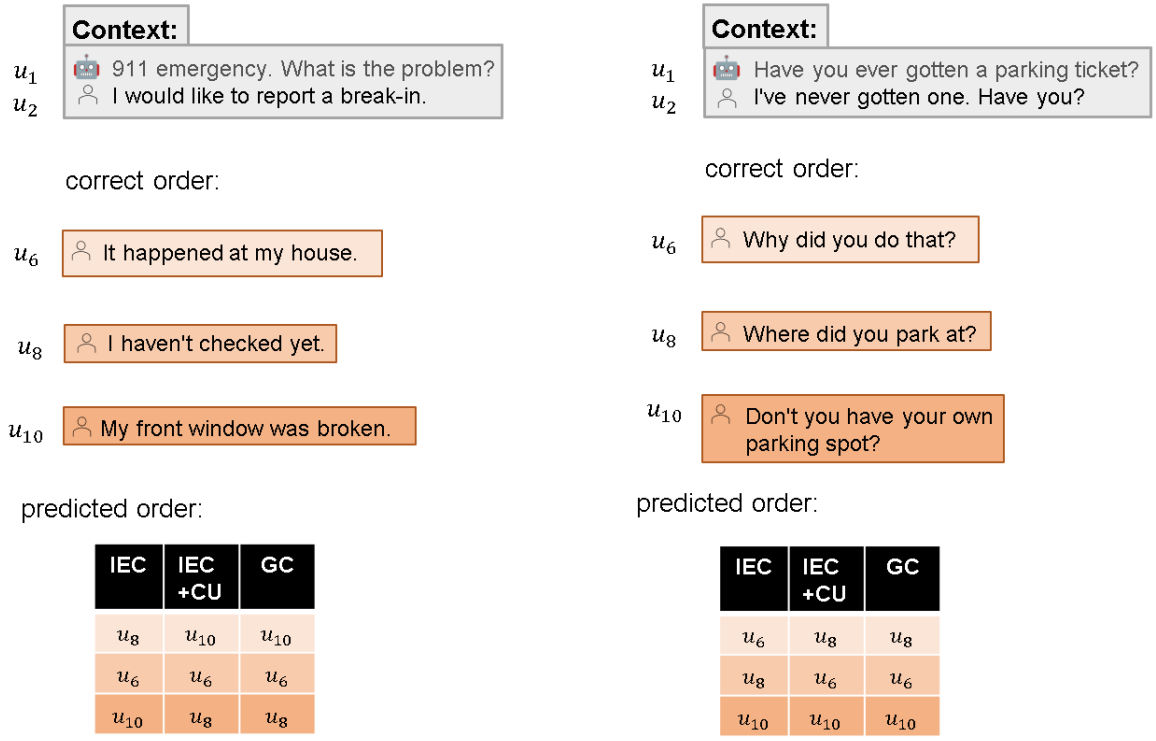

Figure 12: Bad Ranking Examples on DailyDialog Test Corpus with history length of 2, the goal distance of 2, and goal in distance of 3