Space Efficient Context Encoding for Non-Task-Oriented Dialogue Generation with Graph Attention Transformer

Fabian Galetzka[1,2,*], **Jewgeni Rose**[1,3,*†], **David Schlangen**[2] and **Jens Lehmann**[3,4]
[1]Volkswagen Innovation Center, Wolfsburg, Germany
[2]Computational Linguistics, University of Potsdam, Germany
[3]University of Bonn, Germany
[4]Fraunhofer IAIS, Dresden, Germany
{jewgeni.rose,fabian.galetzka}@volkswagen.de
jens.lehmann@iais.fraunhofer.de
david.schlangen@uni-potsdam.de

Abstract

To improve the coherence and knowledge retrieval capabilities of non-task-oriented dialogue systems, recent Transformer-based models aim to integrate fixed background context. This often comes in the form of knowledge graphs, and the integration is done by creating pseudo utterances through paraphrasing knowledge triples, added into the accumulated dialogue context. However, the context length is fixed in these architectures, which restricts how much background or dialogue context can be kept. In this work, we propose a more concise encoding for background context structured in the form of knowledge graphs, by expressing the graph connections through restrictions on the attention weights. The results of our human evaluation show that this encoding reduces space requirements without negative effects on the precision of reproduction of knowledge and perceived consistency. Further, models trained with our proposed context encoding generate dialogues that are judged to be more comprehensive and interesting.

1 Introduction

Building on the idea of attention-based seq2seq models (Vaswani et al., 2017), recent language models such as BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) enable neural conversational models to generate responses that appear human-like and engaging (Yu et al., 2019). A closer look, however, reveals that the lack of long-term memory to represent consistent (world) knowledge and personality over multiple speaker turns can lead to incoherent content being generated (Li et al., 2016; Serban et al., 2017). Initiated by the Conversational Intelligence Challenge (Burtsev et al., 2018; Dinan et al., 2020), the research focus therefore shifted towards *knowledge-grounded* dialogue generation, resulting in first promising approaches using Transformer-based architectures (Dinan et al., 2019; Ghazvininejad et al., 2018; Galetzka et al., 2020).

The basic idea of these approaches is to provide the required background knowledge together with the current dialogue context when decoding the next system utterance. As the underlying language model's input sequence length is limited – for instance, to 1024 tokens in the case of GTP-2 – the presentation of the background knowledge to the model highly impacts the amount of context information that can be fed into a Transformer network. In these earlier attempts, the knowledge was *paraphrased* into pseudo-utterances, on a par with the utterances from the dialogue history. In this paper, we show that a *structured knowledge representation* offers advantages over unstructured text: facts and complex relationships between different entities can be encoded concisely without performance drop in key indicators, such as knowledge correctness, consistency, and interestingness. Chaudhuri et al. (2019) showed the general feasibility of integrating knowledge graphs into domain-specific dialogues. With this work, we integrate arbitrary knowledge graphs into open-domain knowledge-grounded dialogues, preserving the information encoded in their structure.

Space Efficient Context Encoding For our proposed encoding, we generate dialogue-specific local knowledge graphs (*subgraphs* of a background knowledge graph) that capture the information relevant to the dialogue (similar to (Chaudhuri et al., 2021)). We transform these subgraphs into a concise representation that fits the input sequence encoding for the underlying language model (GPT-2): Labels of the *distinct* nodes and edges (entities and corresponding relations) are concatenated with the dialogue history. To preserve the graph structure,

*The first two authors contributed equally to this paper.
†Corresponding author

we fit the attention mask to force the self-attention layers for each node to attend to only connected nodes in the original graph (if there is a connection, attention weight is set to 1, otherwise to 0). This resembles the message-passing approach of graph neural networks (Gilmer et al., 2017).

Naive concatenation of graph triples has a space complexity of $\mathcal{O}(n \cdot k)$, with n being the number of triples and k the number of word tokens per verbalized triple. Paraphrasing these triples into pseudo-utterances results in even larger space complexity. Our proposed encoding has a space complexity of $\mathcal{O}(l)$, with l being the number of distinct node and edge labels (entities and relations). This reduces the required context space compared to triple concatenation or paraphrasing if entities are repeated in the triples (and hence $l < n \cdot k$), which can be assumed to be the case in knowledge graphs (see discussion below). The space savings *grow* with the size and average degree (connectedness) of the graph. Empirical results with two different knowledge-grounded dialogue datasets confirm our theoretical considerations and show that we can reduce the required space by a factor of up to 3.6. These results imply that we can feed more context information into the model, which should result in higher accuracy. We discuss these results in detail in Section 4.3.

Contributions We propose an approach to integrate a concise encoding of knowledge graphs into a Transformer-based decoder architecture for knowledge-grounded dialogue generation. Transformers for natural language generation can be viewed as graph neural networks which use self-attention (Veličković et al., 2018) for neighborhood aggregation on fully-connected word graphs (Xu et al., 2019). We utilize this relationship and restrict the self-attention weights to match the underlying graph structure. Our comprehensive human evaluation with models trained with the publicly available datasets KOMODIS (Galetzka et al., 2020) and OPENDIALKG (Moon et al., 2019), both providing dialogues enriched with structured knowledge, shows that we can reduce the space requirement for context without negative effects on the precision of reproduction of knowledge and perceived consistency. Moreover, our models generate dialogues that are judged to be more detailed and interesting. For reproducibility, we publish all necessary source code and data (https://github.com/fabiangal/space-efficient-context-encoding-acl21).

2 Knowledge-Augmented Neural Conversational Models

Neural conversational models can be categorized into retrieval-based approaches (Lowe et al., 2015; Wu et al., 2017) that choose a next utterance from a set of suitable candidates, and generative approaches (Serban et al., 2016; Wolf et al., 2019; Chaudhuri et al., 2019; Roller et al., 2021) which decode the next utterance token by token out of a fixed vocabulary. The architectures are based on recurrent neural networks such as LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014) cells or self-attention layers (Vaswani et al., 2017) in sequence-to-sequence structures. To integrate knowledge in addition to the dialogue history these models can be augmented by additional recurrent cells to encode the knowledge into a fixed-sized vector representation (Young et al., 2018; Parthasarathi and Pineau, 2020; Ghazvininejad et al., 2018). This can be traced back to first end-to-end approaches reading documents for question-answering (Miller et al., 2016) or more general sequential data (Sukhbaatar et al., 2015). He et al. (2017) embedded knowledge graphs (stored as triples) with LSTM cells and message-passing, and then used a decoder LSTM to generate a suitable answer. Long et al. (2017) used a CNN architecture to encode external knowledge instead.

The recent success of unsupervised pre-trained language generation models such as GPT-2 yielded a variety of conversational models using self-attention based on the idea of fine-tuning the models with specific knowledge-grounded dialogue datasets (which we will discuss in Section 3). These models concatenate the additional context information as plain text to the input sequence (Zhang et al., 2018; Dinan et al., 2019; Galetzka et al., 2020). To differentiate context from dialogue, additional tokens are learned during fine-tuning and added to the word tokens. For bigger knowledge graphs, the limitation of the input sequence length of these models makes an information retrieval system necessary to estimate a small subset of relevant information that can be fed into the model.

3 Knowledge-Grounded Dialogue Datasets

The increasing availability of conversational content on social media platforms such as Twitter or Reddit led to the construction of many dia-

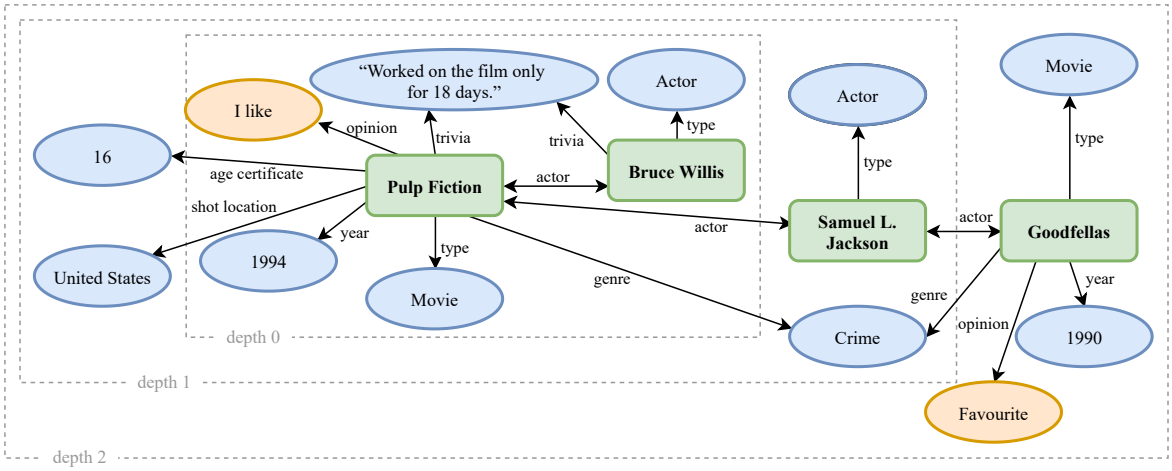

Figure 1: Illustration of the underlying subgraph data model for the external knowledge of a KOMODIS dialogue for different graphs depths: Nodes (green) with their fact-based attributes (blue) and opinions (orange). Subgraphs for depth 1 and depth 2 are incomplete.

logue datasets, with Open-Subtitles (Vinyals and Le, 2015) and Twitter-Corpus (Sordoni et al., 2015) being some popular examples (see also (Ritter et al., 2010; Duplessis et al., 2016)).

Some recently published datasets emphasize knowledgeable dialogues by integrating external information sources. The objective is to create models that generate consistent dialogues with a high knowledge retrieval accuracy (utilizing information from user profiles or knowledge graphs). Dinan et al. (2019) released the Wizard of Wikipedia dataset with over 22k open-domain dialogues. In each dialogue, one participant is playing the "wizard", i.e. an expert who is presented with potentially interesting and relevant Wikipedia article excerpts, while the chat partner is the curious apprentice. The textual knowledge passages that were shown to the wizard are part of the dataset. The PERSONA-CHAT dataset (Zhang et al., 2018) contains over 10k dialogues that are conditioned on profile information (*personas*), which ranges from hobbies or favorite food to family background. The information is shown to the participants as a set of sentences and they are tasked to integrate them into the dialogues. In addition, the dataset contains revised personas, which are rephrased, generalized, or specialized versions of the original personas.

3.1 Dialogue Datasets with Knowledge Graphs

We use two publicly available human/human multi-turn dialogue datasets that use structured background knowledge.

KOMODIS (Galetzka et al., 2020) is a closed-domain dataset with dialogues between human participants that were tasked to chit-chat about one given movie and use provided information about it. This information includes facts about the film, such as release year or shot location ("Movie was shot in Canada." or "The release year is 1995."), free text containing plot or trivia related to the film crew and cast, and opinions towards the facts and entities ("I agree with the age restriction." or "I don't like Bruce Willis."). The dataset contains over 7,500 conversations with an average of 13.8 utterances per dialogue.

OpenDialKG (Moon et al., 2019) is an open-domain dataset containing 15K dialogues, which were collected in a Wizard-of-Oz setup, by connecting two human participants that were tasked to have an engaging dialogue about a given topic. Each dialogue is paired with its corresponding "KG paths" from Freebase (Bollacker et al., 2007) (connecting entities and relations mentioned in the dialogue).

3.2 Subgraph Generation

For our experiments with different encoding strategies, we restructure the context information provided by both datasets into dialogue-specific subgraphs. Figure 1 illustrates an example of an (incomplete) subgraph that belongs to a dialogue from KOMODIS. The inner subgraph containing the two green entity nodes 'Pulp Fiction' and 'Bruce Willis', and corresponding attribute nodes (blue), marked as *depth 0*, represents the information on which one particular dialogue was based.

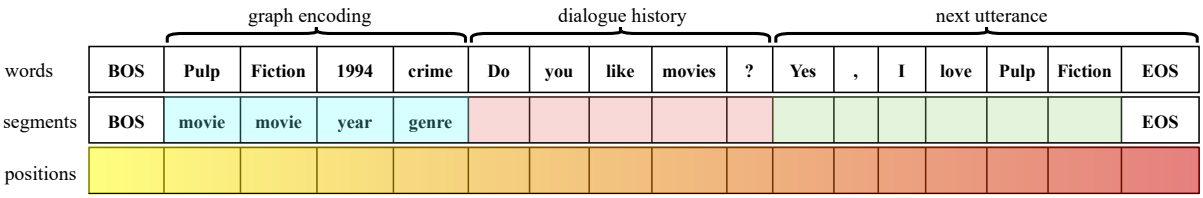

Figure 2: Shortened illustration of the input sequence with encoded context sequence, dialogue history and next utterance with three layers of embeddings: word, segment and positional embeddings. The layers are summed up to yield the by-token embeddings.

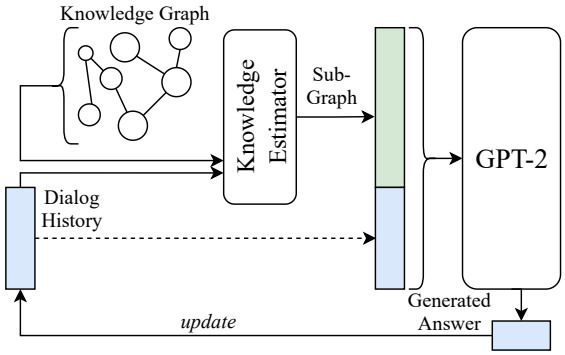

Figure 3: Model architecture: A knowledge estimator creates a subgraph based on the previous conversation. Processed subgraph and input sequence are concatenated and fed into the GPT-2 decoder. We experiment with different ways of encoding and adding in the knowledge.

To test the limits of the capacity for representing knowledge, we also experiment with expanded subgraphs—*depths 1* and *depths 2* in the figure—by including information from external knowledge sources (IMDb for KOMODIS, and Freebase for OPENDIALKG). For instance, Pulp Fiction also has Samuel L. Jackson as an actor (depth 1) who also stars in Goodfellas (depth 2). This way, the subgraph depth directly reflects the hop distance from the entities in the core subgraph.

For subgraphs of depth 2, we restrict some attributes and entities to prevent the subgraphs to explode in size, thus unlikely to fit in GPT-2. For example, we don't add trivia information that isn't already in the dialogues or limit additional actors per movie to three. In contrast to OPENDIALKG, the dialogues in KOMODIS are about one main entity (here, the movie) each. To better compare the experiments across datasets, we create two versions of depth 1 for KOMODIS, where depth 1b includes a second movie that is related to the first movie (e.g. by an actor). This version is then used to create the subgraph of depth 2.

4 Graph Attention Transformer

4.1 Model Overview

For all experiments, we use the GPT-2 model proposed by Radford et al. (2019), which is commonly used in Transformer-based dialogue generation for English. The authors published four different sized variations. We use the model with 117 million parameters, 12 self-attention layers, and 768-dimensional word embeddings. The model has 12 heads per attention layer and 3072 nodes in all feed-forward layers. Our architecture is visualized in Figure 3. A knowledge estimator creates a subgraph from the available knowledge graphs for both datasets based on the dialogue history and converts it using our encoding. Then, the dialogue history and encoded context sequences are concatenated and fed into the GPT-2 model. For training, we optimize model weights from GPT-2 by minimizing the negative log-likelihood for next-token prediction. Training details are listed in Appendix B.

4.2 Concise Graph Encoding

Figure 2 shows the general encoding strategy that we propose. Similar to our previous approach (Galetzka et al., 2020) and Wolf et al. (2019), we use three layers of input embeddings for words, segments and positions. But instead of concatenating paraphrased triples (e.g. ⟨'*Pulp Fiction*', '*is a*', '*movie*'⟩, ⟨'*Pulp Fiction*', '*release year*', '*1994*'⟩), we convert the graph into unique entity-relation pairs (e.g. ⟨'*Pulp Fiction*', '*movie*'⟩, ⟨'*1994*', '*release year*'⟩ in the leftmost part of the figure) and concatenate them with the dialogue history (middle part in figure). In previous work, the segments layer distinguished context and different speakers. We experiment with two different encoding strategies, utilizing the segments layer in other ways. Figure 4 illustrates both encoding strategies. In the series encoding (upper half of the figure), relation and entity tokens are sequenced in a se-

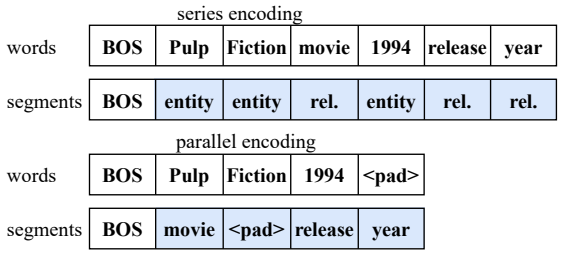

Figure 4: Illustration of the difference between series and parallel encoding with data from the example graph in Figure 1.

ries and added to the words layer. Two new tokens (⟨*entity*⟩ and ⟨*relation*⟩) differentiate between relations and entities in the segments layer. In the parallel encoding, entity tokens are added to the words layer and according relations to the segments layer—thus in *parallel*. Padding tokens are used to align the length between the two layers.

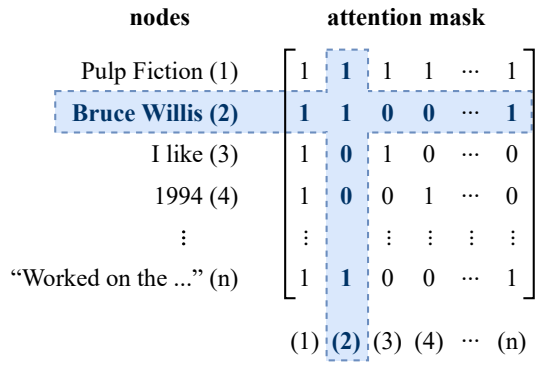

Figure 5: Simplified and shortened illustration of the attention mask for the example graph from Figure 1. The node '*Bruce Willis*' (highlighted in blue) is connected (ones) with the movie '*Pulp Fiction*' and the trivia '*Worked on the ...*'. Other nodes ('*I like*', '*1994*') are masked out (zeros), since they only belong to the movie.

This encoding via a segments layer reduces the space requirements compared to paraphrasing, as repeating tokens occur only once, but on its own loses information encoded in the graph structure (node-edge connections). To preserve this structure information, we create and add a per-graph attention mask to all hidden layers. Given an input sequence S, the hidden state h_i^l of the i'th token at layer l in the GPT-2 model can be computed by:

$$h_i^l = \sum_{j \in S} w_{ij}(V^{l-1}h_j^{l-1}), \tag{1}$$

where

$$w_{ij} = \text{softmax}_j(m_j + Q^{l-1}h_i^{l-1} \cdot K^{l-1}h_j^{l-1}), \tag{2}$$

with learnable weights K, Q, and V. Equation 1 is similar to message-passing algorithms (Duvenaud et al., 2015; Li et al., 2016; Gilmer et al., 2017), where a new hidden state for a graph node is computed by an arbitrary function of all previous hidden states of connected nodes. Our attention masks m_j are added as shown in Equation 2 so that entity and relation tokens can only attend to tokens from their neighboring nodes. This attention masking was originally used for mask out future tokens (setting $m_{i,j}$ for all $j > i$ to the masking value).

Figure 5 illustrates the concept with an attention mask of the graph example from Figure 1. Here, the node '*Bruce Willis*' (blue) is not connected with the release year '*1994*'. Thus, the attention weights are masked out with zeros. But, it is connected with the trivia information '*Worked on the movie for only 18 days*' and *these* attentions are not masked (ones).

Although entities and relations from the knowledge graph are position invariant within S, the word order still matters. Therefore, we keep the positional encoding of the model but shuffle the knowledge graph nodes and relations for each training sample to facilitate order invariance of the graph encoding.

4.3 Context Length Requirement

Figure 6 shows the growth of the number of required context tokens when the graph size is increased (and hence, more knowledge is provided to the model), for different encoding types. The baselines are paraphrased-based encodings, where *base-triples* are the concatenated triples (*"Pulp Fiction release year 1994"*) and *base-paraphrased* the verbalized paraphrase (*"The movie Pulp Fiction was released in 1994"*). For OpenDialKG, no paraphrased version is available. For both datasets, the average number of tokens increases with the graph depth and the average number of nodes and relations for all encodings, as expected. However, it grows much slower in the case of our proposed encodings.

The increase of required tokens for OpenDialKG is steeper than for KOMODIS, due to the different structure of the dialogue context and the underlying knowledge graphs. The context graph for OpenDialKG is initially rather small

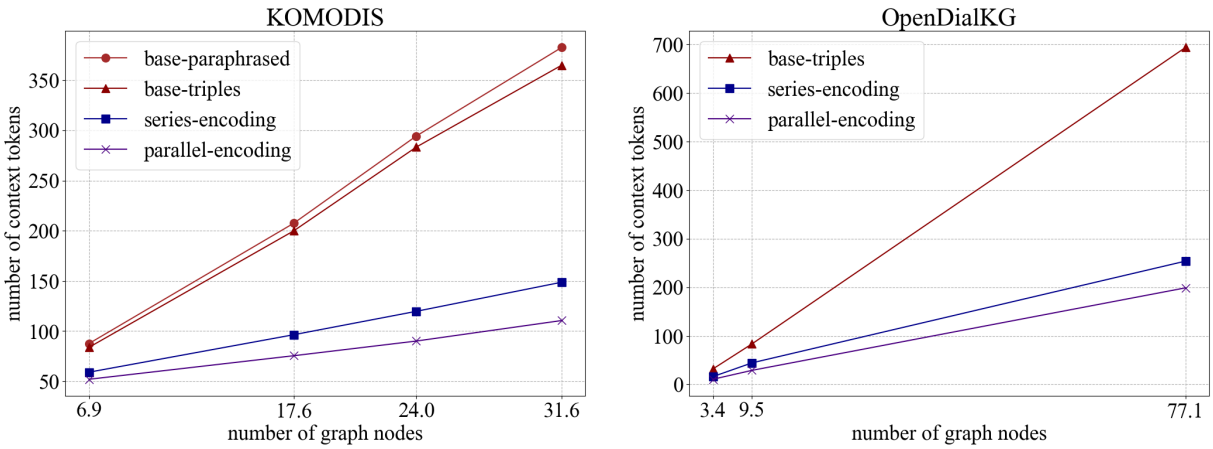

Figure 6: Average number of context tokens in the input sequence for different encodings and knowledge graph depths (KOMODIS from left: d0, d1a, d1b, d2; OPENDIALKG from left: d0, d1, d2). Data extracted from the whole train subset.

and increases very fast with more hops. Further, the KOMODIS context graph contains information about plot and trivia, which are normally longer strings that belong to *one* entity, thus the benefit of series-encoding (*series-enc*) and parallel-encoding (*parallel-enc*) regarding this information is rather small compared to the baselines. Concluding, the sequence length reduction correlates with the average number of edges per node. The *series-enc* is between 14% and 30% longer than the *parallel-enc*, due to representing relation labels within the segments instead of word embeddings (as shown in Figure 4).

5 Automated Evaluation, And Its Limits

We trained 25 models with both datasets with *series-encoding*, *parallel-encoding*, *base-triples* and *base-paraphrased* (only KOMODIS) and with graph depths *d0*, *d1* and *d2*. As we were also interested to investigate the effect of different decoding strategies, we used *beam-search* and *top-k-sampling* when generating the dialogues. These were created by four colleagues (who were not involved in the creation of the models and did not know what the innovation was) interacting with the models. In sum, we created 500 dialogues.

At training time, we use perplexity on the validation subset as the stopping criterion. Table 3 lists the results for all models estimated on the test set. Base-triples (baselines) models reach the lowest perplexity and an increasing graph depth increases perplexity, which is reasonable since the format of the baseline encodings resembles the pre-training data of the GPT-2 model the most. This correlation is stronger for OPENDIALKG models. In our experiments, perplexity does not correlate with naturalness (estimated by the human evaluators), indicating that this metric is insufficient to evaluate the overall dialogue quality, in particular across models trained on different datasets.

In information retrieval, precision/recall and F1 score are typically chosen as automated metrics to evaluate the retrieval capabilities of a system. In our case, we are interested in the ability of reproducing entities and relations from the knowledge graph. Commonly used word-overlap metrics, such as ROUGE-N/ROUGE-L (Lin, 2004), or BLEU (Papineni et al., 2001), aren't capable of measuring these. Evaluating precision and recall requires precise co-reference resolution, parse tree annotation and question labelling (e.g. entities, relations, intent). Assume following background facts and generated utterance:

- ⟨'*Pulp Fiction*', '*has genre*', '*crime*'⟩
- ⟨'*comedy*', '*bot opinion*', '*I like*'⟩
- "*It is a crime movie, but I am more interested in comedy films.*"

Without knowing the intent (asking for the genre of 'Pulp Fiction' vs. asking for the preferred genre) of the previous utterance, we cannot determine if the occurrences of *crime* and *comedy* are true or false positives concerning the underlying knowledge graph. Moreover, in a conversation knowledge is not only used when answering a question

dialogue		context	inference
model	*"have you seen ghost, the musical with brad pitt?"*	genre: drama	neutral
human	*"no, but is the genre really a musical?"*		
model	*"yes, it's a crime and drama movie."*	genre: drama	entailment
human	*"nice. what is the age certificate?"*		
model	*"the age restriction is 12. i think that's a good age to watch it ."*	age certificate: 12 age certificate: agree	entailment entailment
human	*"and do you know the budget?"*		
model	*"yes, it was $24,000,000. i think that's a lot of money for a movie that came out in 1990."*	budget: 22$ million year: 1990	contradiction entailment

Table 1: Example of a dialogue from the series-enc-d1 model trained on KOMODIS. Relevant utterance-context pairs are shown together with annotations from our human evaluators. Line one is labeled with "neutral", as a "drama" might be a "musical", the utterance in line three entails with the context regarding the genre and is therefore labeled with "entailment". Line seven is labeled with "contradiction" as the value of the budget is wrong, but also as "entailment", as the proactive use of the release year is correct.

but also proactively ("*Did you know that Bruce Willis worked on the movie only for 18 days?"*). However, the resources that we use offer no such annotations and to the best of our knowledge, no published dataset does. Without it, automated metrics don't work well. To tackle these challenges, we put our effort into a comprehensive human evaluation and annotation, which is described in the next section.

6 Human Evaluation

6.1 Method

Participants The evaluation study was managed by researchers not involved in setting up the models and experiments. They recruited 20 participants not familiar with our research and the goals of the study. Demographic data is given in Appendix A. Participants were paid for their effort.

Materials To keep the number of assessed dialogues manageable, we limited the number of experiments and did not test all possible variations of the factors described in Section 5. We prepared three series of experiments, aimed at evaluating the influence of *decoding algorithms*, *encoding strategies* and *graph depths*. Early samples indicated that beam-search generates more precise dialogues regarding context. We, therefore, decided to evaluate the decoding algorithm series beforehand. As shown in Section 6.2 our hypothesis proved to be correct, so that the other two series of experiments were done with beam-search only.

Procedure All participants were instructed before and supervised during the study by a supervisor to ensure their understanding of the metrics. They were given a participant-specific questionnaire with the human/chatbot dialogues and had to perform three tasks. First, mark utterances that either entail (correct use) or contradict (wrong use) the dialogue context. Based on these annotations we measure the model's knowledge retrieval ability as the ratio between entailing utterances and the sum of entailing and contradicting utterances (*precision*). Second, rate the dialogues with the following statements for agreement on a 7-point Likert scale: (1) Person B sounds natural. (2) Person B sounds consistent. (3) Person B sounds interesting. Person B is always a model, Person A a human. Last, choose between two dialogues, by answering: "To which Person B would you prefer to talk?". Additionally, the participants could briefly reason their decision. An example questionnaire can be found in Appendix A.

6.2 Results and Discussion

Decoding Table 2 shows the results for beam-search and top-k-sampling decoding. Knowledge precision is better with beam-search for all models, while dialogues generated with top-k-sampling are considered more natural, less self-contradicting, and less repetitive. N-gram filtering reduces repetition through beam-search, but could not be avoided completely. Decoding with top-k-sampling includes more often wrong entity nouns when es-

experiment	knowledge precision		naturalness	
	base-triples	series-enc-d1	base-triples	series-enc-d1
KOMODIS beam-search	**0.69**	**0.74**	5.0 (1.5)	4.8 (1.6)
KOMODIS top-k-sampling	0.52	0.56	**5.9** (1.2)	**5.9** (1.3)
OPENDIALKG beam-search	**0.73**	**0.70**	4.0 (1.6)	3.4 (1.5)
OPENDIALKG top-k-sampling	0.54	0.45	**5.3** (1.4)	**5.4** (1.3)

Table 2: Human evaluation results for beam-search and top-k-sampling, with respect to the correct reproduction of dialogue context. Precision as the ratio between entailing utterances and the sum of entailing and contradicting utterances. Naturalness on a 7-point Likert scale. Higher is better. Standard deviation in brackets.

	experiment	ppl	win ratio (%)	precision		agreements		
				knowledge	opinions	natural	consistent	interesting
KOMODIS	base-paraphrased	10.3	12.5	**0.74**	0.50	4.7 (1.7)	4.1 (1.7)	4.2 (1.2)
	base-triples	**9.73**	43.8	0.69	**0.71**	**5.0 (1.5)**	4.0 (2.0)	4.6 (1.1)
	series-enc-d1	10.01	**66.7**	**0.74**	0.36	4.8 (1.7)	4.5 (1.9)	**4.9 (1.1)**
	series-enc-d2	10.28	62.5	0.73	0.43	4.8 (1.7)	4.2 (1.6)	4.4 (1.2)
	parallel-enc-d1	10.07	56.3	0.70	0.33	4.5 (1.7)	4.5 (1.2)	4.5 (1.1)
	parallel-enc-d2	10.36	60.0	0.72	0.57	4.8 (1.5)	**4.6 (1.5)**	4.5 (1.2)
OpenDialKG	base-triples	**8.40**	65.0	**0.73**	—	**4.0 (1.6)**	3.9 (1.6)	3.6 (1.6)
	series-enc-d1	9.93	**66.7**	0.62	—	3.9 (1.9)	**4.1 (1.9)**	3.5 (1.9)
	series-enc-d2	10.53	51.3	0.46	—	3.7 (1.7)	4.0 (2.0)	**3.8 (1.6)**
	parallel-enc-d1	9.88	38.5	0.70	—	3.4 (1.6)	3.2 (1.8)	3.0 (1.3)
	parallel-enc-d2	10.44	32.5	0.62	—	3.4 (1.9)	3.6 (1.9)	3.3 (1.6)

Table 3: Perplexity on the test set (lower is better) and human evaluation results for models trained on both datasets. Metrics explained in Section 6.1. Agreements are on a 7-point Likert scale (higher is better). Standard deviation in brackets. "base-*" are the baseline models; "series/parallel-enc-*" denotes the way the knowledge is encoded and "-*d1/d2" is the depth of the graphs.

timating the best next tokens, which are then selected by the algorithm. In this work, we emphasize the model's ability to integrate additional dialogue context correctly. Here, models with beam-search perform significantly better. Thus, our further evaluation focuses on beam-search.

Graph Encoding The results with series and parallel graph encodings are shown in Table 3 and compared against the baselines. Within each dataset, all models perform similar regarding knowledge precision. Due to the high standard deviation on the agreements, the difference between the models is statistically insignificant. Our graph encoding approach reduces the required input sequence length by a factor of up to 3.6 and still achieves the same quality of knowledge reproduction, consistency, and naturalness as the baselines. Further, the direct dialogue comparison (win ratio) indicates more comprehensive and interesting utterances for KOMODIS. Dialogue preference correlates highest with interestingness and non-existence of contradicting statements. The most common reasons from participants in no specific order are "longer and more comprehensive utterances", "more interesting", "asks counter questions" and "more pleasant". The OPENDIALKG models perform worse in general but show similar results between the different encodings. Both datasets have similar sizes but OPENDIALKG is not limited to the movie domain, which makes it harder to train compared to KOMODIS.

Series vs. Parallel Encoding A quick summary: the segments layer encodes the typing of the word tokens (from the words layer). The intuition behind it is that the model learns the *meaning* of the words instead of the word distribution alone. For the series encoding, we encode the types generically as either entity or relation. For the parallel encoding, we use the actual typing from the underlying knowledge graph, such as movie, actor, or release year (Section 4.2). We had two objectives. First, reducing the required context space even fur-

ther (which we achieved, see Figure 6). Second, analyzing if this improves the accuracy. The results show, that parallel encoding performs slightly worse compared to series encoding. We assume that this is the case due to the lack of training data, which is, in particular, evident for OPENDIALKG that has much more entity and relation types than KOMODIS, i.e. fewer samples per type.

Graph Depth Results for training with different context lengths with KOMODIS are shown in Table 4. All metrics (one outlier for opinion precision with $d = 1$) correlate with increasing graph depth. Results for $d = 2$, however, are statistically not significantly higher than for $d = 1$. A bigger subgraph leads to more difficult training data, as the model has more options to choose from. The same results couldn't be reproduced for OPENDIALKG. This dataset was created for graph generation based on dialogues. However, the dialogue structure is different due to the recommendation task of the data collection. Most entities in these dialogues (e.g. persons, books, movies) are exchangeable ("Can you recommend me a crime book similar to X?", "Can you recommend me a crime movie similar to Y?") and therefore not mandatory for a correct and consistent dialogue. Adding more of these entities did not help to determine a correct next entity, as all entities of the same type could be used correctly by the model.

Effectiveness of Graph Attention Masking Graph masking encodes the relationships between the entities. We hypothesize that dropping these relationships will lead to an information gap, particularly for bigger subgraphs due to more entities that are not represented (well) in the training data. Table 5 shows the results from an early evaluation phase for KOMODIS and OPENDIALKG with graph depth 1 and 2 without graph masking. The dialogues are significantly worse, in particular in terms of reproducing entities correctly for graph depth 2 – which validates our hypothesis. As our resources were limited, we had to reduce the number of models for a thorough human evaluation and thus decided to not pursue this approach any longer.

7 Conclusion

We proposed a new and concise encoding for knowledge triples from a knowledge graph, which can be integrated into a Transformer architecture for consistent non-goal-driven dialogue generation. In our encoding, we reduce the context length by avoiding repetition by concatenating the whole triples with the dialogue history. By manipulating self-attention layers to reflect connections between nodes in the graphs, we preserve the graph structure. The evaluation results prove that our encoding reduces space requirements without negative effects on the precision of reproduction of knowledge and perceived consistency. For reproducibility, we publish the source code and data.

metric	d0	d1	d2
knowledge precision	0.56	0.70	**0.72**[1]
opinion precision	0.42	0.33	**0.57**
naturalness	4.5	4.5	**4.8**[1]
win-ratio (%)	28.6	56.3	**60.0**[1]

Table 4: Influence of graph depth on various metrics from the human evaluation for the parallel-enc model trained on KOMODIS. [1]Statistically not significant compared to d1.

experiment	knowledge	opinions	naturalness
KOMODIS d2	0.44	0.25	4.4
KOMODIS d1	0.61	0.46	4.1
OPENDIALKG d2	0.37	—	3.8
OPENDIALKG d1	0.54	—	3.9

Table 5: Results from a pre-evaluation for models without graph attention masking. There are no opinions in the case of OPENDIALKG. Knowledge and opinions as precision (ratio between entailing utterances and the sum of entailing and contradicting utterances). Naturalness on a 7-point Likert scale. Higher is better. Standard deviation in brackets.

Acknowledgements

We thank our colleagues from the Digital Assistant for Mobility team at the Volkswagen Group Innovation Europe for their support in preparing the human evaluation.

References

Kurt Bollacker, Robert Cook, and Patrick Tufts. 2007. Freebase: A shared database of structured general human knowledge. *Proceedings of the national conference on Artificial Intelligence*, 22(2):1962.

Mikhail Burtsev, Varvara Logacheva, Valentin Malykh, Iulian Vlad Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, and

Yoshua Bengio. 2018. The first conversational intelligence challenge. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 25–46. Springer.

Debanjan Chaudhuri, Md Rashad Al Hasan Rony, Simon Jordan, and Jens Lehmann. 2019. Using a KG-Copy Network for Non-goal Oriented Dialogues. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11778 LNCS:93–109.

Debanjan Chaudhuri, Md Rashad Al Hasan Rony, and Jens Lehmann. 2021. Grounding Dialogue Systems via Knowledge Graph Aware Decoding with Pre-trained Transformers. pages 1–16.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1724–1734.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 4171–4186.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, Jason Weston, and Others. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Of Wikipedia: Knowledge-powered conversational agents. *7th International Conference on Learning Representations, ICLR 2019*, pages 1–18.

Guillaume Dubuisson Duplessis, Vincent Letard, Anne Laure Ligozat, and Sophie Rosset. 2016. Purely corpus-based automatic conversation authoring. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 2728–2735.

David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems*, 2015-Janua:2224–2232.

Fabian Galetzka, Chukwuemeka Uchenna Eneh, and David Schlangen. 2020. A Corpus of Controlled Opinionated and Knowledgeable Movie Discussions for Training Neural Conversation Models. *Proceedings of The 12th Language Resources and Evaluation Conference*, (May):565–573.

Marjan Ghazvininejad, Chris Brockett, Ming Wei Chang, Bill Dolan, Jianfeng Gao, Wen Tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 5110–5117.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. *34th International Conference on Machine Learning, ICML 2017*, 3:2053–2070.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:1766–1776.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. 2016. Gated graph sequence neural networks. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, (1):1–20.

C Y Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Proceedings of the workshop on text summarization branches out (WAS 2004)*.

Yinong Long, Jianan Wang, Zhen Xu, Zongsheng Wang, Baoxun Wang, and Zhuoran Wang. 2017. A Knowledge Enhanced Generative Conversational Service Agent. *DSTC6 Conference*, pages 1–5.

Ryan Lowe, Nissan Pow, Iulian V. Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A large dataset for research in unstructured multi-turn Dialogue systems. *SIGDIAL 2015 - 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, pages 285–294.

Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 1400–1409.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Acl*, volume 371, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Prasanna Parthasarathi and Joelle Pineau. 2020. Extending neural generative conversational model using external knowledge sources. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 690–695.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, (June):172–180.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-To-end dialogue systems using generative hierarchical neural network models. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pages 3776–3783.

Iulian Vlad Serban, Alessandro Sordoni, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. *Aaai2017*, pages 3295–3301.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 196–205.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. *Advances in Neural Information Processing Systems*, 2015-Janua:2440–2448.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5999–6009.

Petar Veličković, Arantxa Casanova, Pietro Liò, Guillem Cucurull, Adriana Romero, and Yoshua Bengio. 2018. Graph attention networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pages 1–12.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. 37.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. (ii).

Yu Wu, Wei Wu, Chen Xing, Zhoujun Li, and Ming Zhou. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:496–505.

Peng Xu, Chaitanya K. Joshi, and Xavier Bresson. 2019. Multi-Graph Transformer for Free-Hand Sketch Recognition.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 4970–4977.

Dian Yu, Michelle Cohn, Yi Mang Yang, Chun Yen Chen, Weiming Wen, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Giritheja Sreenivasulu, Sam Davidson, Ashwin Bhandare, and Zhou Yu. 2019. Gunrock: A Social Bot for Complex and Engaging Long 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): System Demonstrations*, pages 79–84, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:2204–2213.

A Human Evaluation

Demographic data 45% of the 20 participants are women. 75% of the participants stated that they already have experience with various forms of chatbots. Due to data privacy reasons, age information is classified into three different categories. 65% of the participants are 18–35 years old, 20% 36–50 years, and three participants are older than 50 years.

Questionnaire The questionnaire contains a survey guide and a set of dialogue pairs to evaluate. An example dialogue pair is shown in Figure 7. Labels were added by the authors. The survey guide consists of four pages with examples and explanations for the participants. The following excerpts are from the guide.

General instructions: *Following, you are presented with two dialogues between Person A and Person B with according background information. The dialogues are completely independent of each other. You must read both dialogues carefully. Please take time for this task.*

Instructions for evaluating the knowledge and opinion precision: *Please remember that the evaluation is for Person B only! Please add 'entailment' to the fields, when the utterance entails a specific fact or opinion. Please add 'contradiction' if an utterance contradicts a specific fact or opinion. Please leave all other fields empty.*

Instructions for rating the dialogues on the 7-point Likert scale: *Please rate the three statements for each dialogue on a scale from 1 to 7, where 1 means that you strongly disagree with it and 7 that you strongly agree with it. Please rate all statements independently from the given facts and opinions. For instance, if a dialogue contains wrong facts, it still can sound very natural.*

Instructions for deciding between two dialogues: *Please rate intuitively with which Person B you would prefer to talk. Please reason your decision briefly.*

All instructions are provided with examples.

B Training Details

For fine-tuning GPT-2, we reused most training parameters from the generative pre-training (Radford et al., 2019). The learning rate linearly decreases to zero with an initial value of $lr = 6.25\text{e}{-6}$ with max-norm gradient clipping. The language modeling loss is multiplied by 2 before summed up with the next sentence classifier loss. Each mini-batch consists of 32 sequences of up to 256 tokens (padded to maximum length). If dialogue history exceeds maximum sequence length, the first utterances are cut off. For each sample, only the tokens from the last utterance are considered for the language modeling loss. Encoded nodes and edges are shuffled randomly for each sample, not for each dialogue. We used a cluster of 4 GeForce RTX 2080 Titan to train our models with batch distribution and gradient accumulation to handle the mini-batch size. Based on graph depth an epoch took up to 4 hours. We trained the models for 7 to 10 epochs. Our graph encoding approaches took longer to converge, compared to the baselines.

C Dialogue Examples

We show three additional typical dialogue examples for both models in Tables 6, 7 and 8. They contain entailing and contradicting statements regarding the context. The inconsistent opinion in Table 8 (correct in line 4, but incorrect in line 6) was observed more often. For both datasets, we sometimes observe slight misspellings of entities (see Table 8, line 6, '*Charlie Chaplin*' and '*Charles Chaplin*'). We left the decision, whether this is a contradiction (a wrong entity) or not (a misspelling) to the participants. Another issue that occurred multiple times, is the wrong use of repeated entities (see Figure 7 with '*Julia Roberts*' in line 6). Models trained with OPENDIALKG also sometimes use entities that are not in the knowledge graph (see Table 7, line 4 '*The Player*').

			Dialogue	Sarfarosh is of genre Action.	Smita Jaykar has trivia: Often plays the role of the mother in Bollywood movies.	Sonali Bendre is an actor in Sarfarosh.	Sarfarosh was released in 1999.	Action is not a good genre.	Sonali Bendre is very good.	I like Sarfarosh very much.	With which Person B would you prefer to talk?	Why?	Person B sounds natural	Person B sounds consistent	Person B sounds interesting
		A	hey, how are you? I watched Sarfarosh yesterday.							contradiction					
		B	i'm good thanks... yourself?. i don't know that movie actually... what's it about?												
		A	I am good! It's an action movie with sonali bendre.						entailment	entailment					
	1	B	oh okay cool... sounds like it might be a good one to watch with your kids.										5	3	4
		A	why do you think so?	entailment				entailment		contradiction					
		B	i don't really care for action movies... so maybe that's why i didn't care for this one.												
		A	what do you like?					entailment							
		B	i just don't think it's a good genre to be a part of... i don't like action movies either.								2	It would be more interesting to talk to the second Person.			
				Bill Raymond is an actor in Twelve Monkeys.	Twelve Monkeys is of genre Sci-Fi.	Twelve Monkeys was released in 1995.	Twelve Monkeys is of genre Mystery.	Age restriction: I agree.	I like Bill Raymond very much.	'Twelve Monkeys' is good.					
		A	hey, how are you?												
		B	i'm good. i just watched a movie called "twelve monkeys". have you seen it?												
		A	yes, but I didn't like it.							entailment					
		B	really? i thought it was pretty good.												
	2	A	Bill Raymond is such a good actor!	entailment					contradiction				6	4	6
		B	i don't really like bill raymond, but he did a good job in this movie.												
		A	Do you know the release year?			entailment									
		B	i believe it was 1995. do you know what the budget was for this movie by any chance?												

Figure 7: Example of one dialogue pair from the questionnaire.

dialogue		context	inference
human	*"I am a fan of FC Bayern Munich!"*		
model	*"they are a good team. they are in the allianz arena right now and they are playing in the bundesliga, do you like them?"*	building: allianz arena team: fc bayern munich	neutral entailment
human	*"Yes I do, what is the Allianz arena?"*		
model	*"it is a building of the fc bayern munich and was built by mark van bommel."*	building: allianz arena player: mark van bommel	entailment contradiction
human	*"Did they participate in the UEFA champions league?"*		
model	*"yes, they did. it is the team that won the 2012-13 champions league."*	champions league 2012-13 champion: fc bayern munich	entailment

Table 6: Example of a dialogue from the series-enc-d1 model trained on OPENDIALKG. Relevant utterance-context pairs are shown with gathered labels from our human evaluation.

dialogue		context	inference
human	*"Do you recommend Pretty Woman?"*		
model	*"pretty woman is starring hank azaria. do you like him?"*	has actor: hank azaria	entailment
human	*"Yes, who else plays in the movie?"*		
model	*"julia roberts is also in pretty woman. she is married to daniel moder, who also starred in the player and a few others."*	has actor: julia roberts j. roberts spouse: daniel moder	entailment entailment
human	*"do they have children?"*		
model	*"yes, they have a daughter named julia roberts."*	j. roberts has child: henry	contradiction

Table 7: Example of a dialogue from the series-enc-d1 model trained on OPENDIALKG. Relevant utterance-context pairs are shown with gathered labels from our human evaluation.

dialogue		context	inference
human	*"Hey, what's up?"*		
model	*"not much. i just watched a movie called city lights, have you seen it?"*		
human	*"No, what genre is it?"*		
model	*"it is a drama and family movie. i really like drama movies, but this one didn't do it for me."*	genre: drama City Lights: don't like	entailment entailment
human	*"So you wouldn't recommend it?"*		
model	*"i think charlie chaplin is a bad director, but he did a good job with this movie."*	director: Charles Chaplin Charles Chaplin: don't like City Lights: don't like	entailment entailment contradiction
human	*"What else do you know about him?"*		
model	*"i don't know much about him, but he cofounded the distribution company united artists in 1919."*	Charles Chaplin has trivia: Co-founded the ... in 1919.	entailment

Table 8: Example of a dialogue from the series-enc-d1 model trained on KOMODIS. Relevant utterance-context pairs are shown with gathered labels from our human evaluation.