Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 565–573
Marseille, 11–16 May 2020

A Corpus of Controlled Opinionated and Knowledgeable Movie Discussions for Training Neural Conversation Models

Fabian Galetzka[1,2], Chukwuemeka U. Eneh[2], David Schlangen[1]
[1]Computational Linguistics, University of Potsdam, Germany
[2]Volkswagen AG
fabian.galetzka@volkswagen.de

Abstract

Fully data driven Chatbots for non-goal oriented dialogues are known to suffer from inconsistent behaviour across their turns, stemming from a general difficulty in controlling parameters like their assumed background personality and knowledge of facts. One reason for this is the relative lack of labeled data from which personality consistency and fact usage could be learned together with dialogue behaviour. To address this, we introduce a new labeled dialogue dataset in the domain of movie discussions, where every dialogue is based on pre-specified facts and opinions. We thoroughly validate the collected dialogue for adherence of the participants to their given fact and opinion profile, and find that the general quality in this respect is high. This process also gives us an additional layer of annotation that is potentially useful for training models. We introduce as a baseline an end-to-end trained self-attention decoder model trained on this data and show that it is able to generate opinionated responses that are judged to be natural and knowledgeable and show attentiveness.

Keywords: Non-Goal Driven Dialogues, Opinionated Discussions, Deep Neural Networks

1. Introduction

Where dialogue modelling used to be mostly rule-based with the dialogue being driven by pre-specified knowledge representations (e.g., (Bobrow et al., 1977), (Traum and Larsson, 2003), (Stede and Schlangen, 2004)), recent years have seen efforts of basing this task on models directly learned from data. A particular strand of this research has modelled the task of producing a dialogue contribution in analogy to the translation task as one of going from one sequence (the user utterance) to another sequence (the system utterance).

The first such models solely based on data driven end-to-end approaches (Li et al., 2016; Serban et al., 2017) tended to generate universal and inconsistent utterances regarding content and personality. We illustrate this problem with the example in Figure 1, distinguishing the two consistency dimensions *knowledge* (a speaker should not "forget" previously known facts) and *opinion* (a speaker should not change their opinion, at least not without any overt trigger in the conversation). In this example, each system response is locally coherent (a good reply to its immediate precursor), but globally inconsistent.

While this particular example is constructed, it is not very far from what these early models would have been liable to produce. One reason for this is that these models were optimised only for local coherence, and trained from datasets such as TwitterCorpus (Ritter et al., 2010) and OpenSubtitles corpus (Tiedemann, 2012). These datasets contain dialogues from many people, without any information about the speakers and their opinions or knowledge state.

To tackle issues like these, several augmented dialogue datasets have been introduced in recent years. Zhou et al. (2018) created a dataset with conversations based on Wikipedia articles about popular movies. Another more general dataset (Dinan et al., 2018) explicitly tasked one person in each conversation to link the used knowledge to each written utterance. Models trained on these augmented datasets produced to more engaging and more natural dialogues, as shown in that paper. As opposed to additional general knowledge, the PERSONA-CHAT dialogue corpus (Zhang et al., 2018) is based on personality profiles. Crowd workers were matched together in a role-playing chat and asked to get to know each other, considering profile information which was individually provided for every participant. Different types of neural networks were trained on that dataset, which were shown to also produce more engaging and consistent dialogues compared to models trained on other datasets.

U1: "Do you know Pulp Fiction?"
S1: "Yes, I love Pulp Fiction."
U2: "Yeah, Pulp Fiction is great."
S2: "I hate that movie."
U3: "You hate Pulp Fiction?"
S3: "I don't know it, is it good?"

Figure 1: A constructed example of a dialogue that is *locally* coherent, but *globally* incoherent along the dimensions *knowledge* (S3 to S1, S2) and *opinion* (S2 to S1).

We contribute to this research a corpus that combines these strands, as it consists of dialogues that were collected in a setting where we controlled both the knowledge available to the dialogue participants, as well as their stance towards entities to be mentioned in it. We present this corpus in the next section, and then show that from it models can be induced that better exhibit global consistency, along these dimensions.

Profile (Person A)	**Profile (Person B)**
Talk about <*movie*>!	Talk about <*movie*>!
Plot of / Fact about <*movie*> *	Plot of / Fact about <*movie*> *
Fact about <*other_entity*> **	Fact about <*other_entity*> **
Opinion about <*movie*>	Opinion about <*movie*>
Opinion about <*other_entity*> **	Opinion about <*other_entity*> **
Task: Ask for specific fact *	

Figure 2: An abstract profile pair $P_{\mathrm{A}}, P_{\mathrm{B}}$, each given to one chat partner. * denotes information that may not occur in every profile. ** denotes the possibility of multiple occurrences with different entities.

2. The KOMODIS-Dataset: Collection

We introduce a new augmented dialogue dataset (**K**nowledgable and **O**pinionated **MO**vie **DIS**cussions) that is crowd-sourced and collected with Amazon Mechanical Turk (AMT). Every dialogue is constrained by a unique set of facts as well as a suitable set of opinions about the entities in the facts. The dataset is set in the domain of movies, which we selected because it is a popular topic likely to generate engaging conversations. The creation of the dataset is described in the present section, its validation relative to the aims of controlling opinion background and knowledgeability is described in detail in Section 3.2. The dataset is publicly available in our online repository[1].

Inspired by (Zhang et al., 2018) our dialogues are collected by providing additional information to the participants. Unlike in that work, however, we do not indicate a textually-described personality, but rather provide facts about entities (knowledge) and opinions about them. For each dialogue we created a unique set of two *profiles* (formalised here as feature structures). In all cases both crowd-worker had to talk about the same movie, with different combinations of feature structures $P_{\mathrm{A}}, P_{\mathrm{B}}$. An abstract example is shown in figure 2. The facts are explained in more detail in 2.1., as well as the opinions in 2.2. The different combinations of feature structures are explained in 2.3. A concrete example is shown in Figure 3.

2.1. Facts

The facts are a combination of three different types of information, all extracted from the publicly available movie database IMDb[2]:

(1) Open-domain sentences, so called trivia about movies and actors. For example: '*The screenplay says that Zed and Maynard are brothers.*', or: '*Quentin Tarantino was quoted as saying that Butch is responsible for keying Vincent's car.*'. These trivia information is itself crowd-sourced in IMDb, but comes with a crowd-sourced rating. We only use such trivia marked as *interesting* in the IMDb. We also used the overall length of the trivia, with shorter trivias preferred over longer ones, to ensure a compact set of facts in the end.

(2) A short plot of every movie. For example from Pulp Fiction: '*The lives of two mob hitmen, a boxer, a gangster's wife, and a pair of diner bandits intertwine in four tales of violence and redemption.*'

(3) Facts like release date or budget of a movie. While the trivia have the form of open-domain sentences, these facts are given as knowledge triples in the database. We created multiple sentence patterns per type of fact to convert them into sentences as well.

Given a specific movie, we took 2–4 facts to generate a set. The facts were chosen randomly with a few constraints to ensure a fluent dialogue. For example, if a randomly selected trivia about a movie mentioned an actor, the next fact could be about that actor and so on:

(1) Sometimes one participant is asked to pretend not to know a certain movie, in which case they do not get any information about it. Instead we provide at least one question.

(2) If one participant gets the task to ask a specific question, we provide the correct answer to the other participant.

(3) We prioritized trivias that include entities of actors from the given movie. If that is the case, we provided additional information about this actor.

(4) We randomly added additional facts like budget or genre, but not every set of facts has one of these information.

(5) Every trivia is only used once in the whole dataset.

2.2. Opinions

We augmented the facts by a set of suitable opinions. For example, if a trivia is about an actor, we provided an opinion about that actor. We used a discrete set of opinions ranging from *really don't like* to *favorite* as well as *don't know*. The attitudes were converted into sentences, too. Their strength was generated randomly and all possible combinations are available.

2.3. Relations between Speaker Profiles

To induce interesting dialogues, we varied the relations between the profiles. In the first type of relation, both have the same profile (knowledge of facts and opinions about them):

$$P_{\mathrm{A}} = P_{\mathrm{B}} \tag{1}$$

We also create profile sets where the individual profiles are complimentary, but not conflicting (e.g., A knows some-

[1] https://github.com/fabiangal/komodis-dataset

[2] https://www.imdb.com/

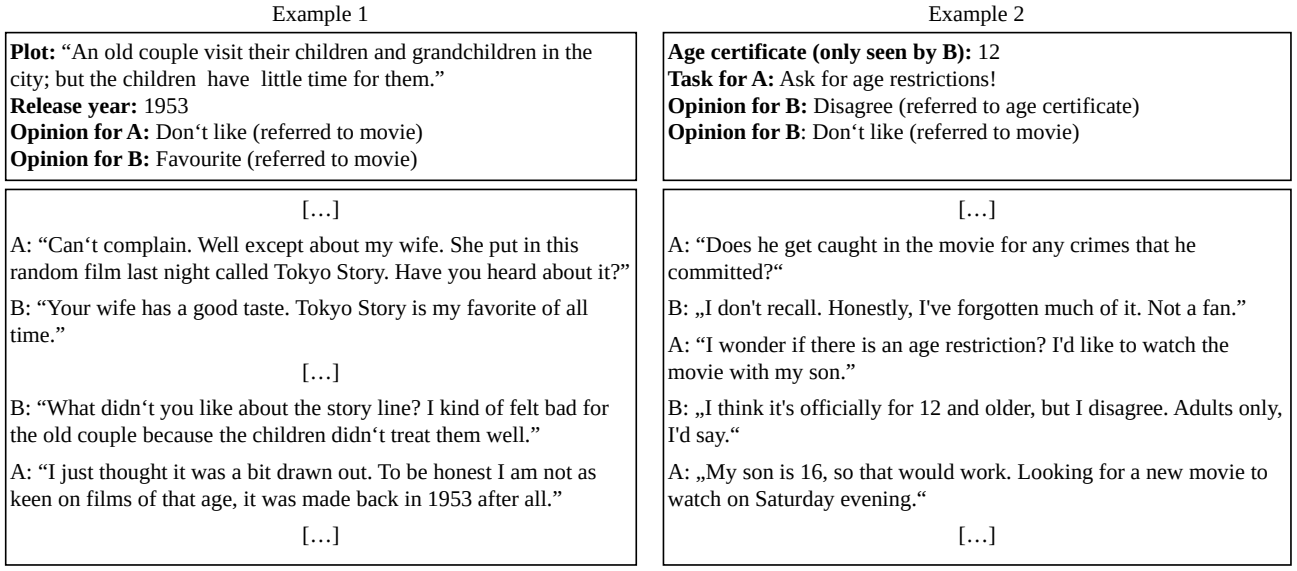

Figure 3: Parts of two dialogues from our corpus with the given relevant facts and attitudes. One can see that both participants combined the given information within the conversation with success. And even though the topic is constrained, the dialogues have open chit-chat characteristics.

thing that B doesn't; formally, the feature structures representing the profiles can be unified):

$$P_{\mathrm{A}} \sqcup P_{\mathrm{B}} \tag{2}$$

Finally, we also created sets with incompatibilities (only along opinions, however, since we did not want them to get into factual discussions):

$$P_{\mathrm{A}} \sqcup P_{\mathrm{B}} = \bot \tag{3}$$

2.4. Collection Procedure

The dataset was collected via Amazon Mechanical Turk and with the help of *slurk*, a chat server developed for data collection tasks that require a matching of multiple participants (Schlangen et al., 2018). Two crowd-worker were paired and tasked to chat using the provided information (different for each of them). The crowd-worker were tasked to use both the facts and opinions to generate sensible utterances within the conversation.

AMT provides two types of payments. A basic one, which is fixed and bound to the task and a flexible bonus payment, that can be paid manually afterwards. Matching two participants requires at least one of them to wait for a partner. We used that process of waiting for the small basic payment. Then, after a successful match, we paid most of the fee for up to three minutes of chatting as bonus payment. If crowd-worker waits for three minutes, they can finish the task without chatting; this happened in less then 5% of the cases though.

In our first iterations we figured out that the crowd-worker tended to simply copy the trivia and rush through the facts. Another problem with written chat is that cross talk can occur (where both participants are typing at the same time and messages get out of order). We found that by enforcing who started the conversation, by giving one randomly selected participant the first turn, we could reduce this, without having to enforce strict turn taking throughout the interaction. This increased the data quality considerably. Also, the quality of the dialogues increased with the amount of money we paid. A bonus payment for 'well written utterances' also helped. We paid up to 1.60 per dialogue. Additionally we limited the number of tasks one crowd-worker can do per day by 5.

After the chat we asked the participants to rate their partner in terms of language quality, naturalness and attentiveness.[3] We speculated that this information might be useful to detect bad quality dialogues, and could also serve as a baseline for human evaluation of trained models.

3. Dataset Overview and Validation

In the following section we present a quantitative overview of our dataset, as well as a detailed validation of the data.

3.1. Dataset Statistics

We initially collected $8,000$ interactions. From these, we had to filter out 1,032 (12.9%) because either one participant did not respond in the conversation or one participant rated his partner's quality negatively. In a second iteration we collected another batch, bringing the total up to $7,519$ dialogues. In these, there is an average number of 13.8 speaker turns ($103,500$ in total). We have split our dataset into a train, validation and test set with 80%, 10% and 10% dialogues respectively, in such a way to no movie is in more than one split. We give some descriptive statistics about the dataset in table 1.

[3] Actual statements the participants had to rate: "My partner is a native speaker of English","This felt like a natural chat about movies","My partner chatted like an attentive person would"

Parameter	Value
dialogues	7,519
utterances	103,500
tokens	1,487,284
average utterances per dialogue	13.8
average tokens per utterance	14.4
vocabulary size	27,658
vocabulary size (99% of tokens)	13,727
different movies	500
used (unique) trivia	13,818
participants from AMT	569

Table 1: Quantitative representation of our dataset. Trivias and movies are approximately evenly distributed over all dialogues. The vocabulary size was computed separately by counting all different tokens from the original dialogues. However, for training we used byte pair encoding.

3.2. Dataset Validation

After collecting the dialogues we post-processed and validated the dataset. As it is not possible to supervise the crowd-worker automatically while chatting, we have to be sure that a) they really talked about the profile entities and b) adhere to the opinions specified there.

3.2.1. Named Entity Resolution

As a first step we extracted all named entities from each dialogue. Even though with the existence of powerful natural language processing tools like Spacy (Honnibal and Montani, 2017) and CoreNLP (Manning et al., 2014), which can detect mentions of names, organizations or countries with high precision (named entitiy recognition, NER), detecting movie titles still remains a challenging problem (Ashwini and Choi, 2014), especially with grammatical errors and spelling mistakes. However, for each dialogue, we knew which movie they were (supposed to be) chatting about, which reduces the complexity of named entity recognition in our domain. We used three different metrics to find an entity: First, exact string match on the lowercased strings, which has high precision but very low recall. Second, we compared every n-gram in an utterance and the movie title with the cosine similarity from Spacy. We used a threshold of 0.9 and

$$\min\left(t_{\text{movie}}; 3\right) \leq n \leq t_{\text{movie}} \tag{4}$$

for the n-grams, with t_{movie} as the number of tokens of a movie title. And third, a combination of the Jaccard distance (Niwattanakul et al., 2013) with threshold ≤ 3 and Levenshtein distance (Levenshtein, 1966) with threshold ≤ 0.3 for the same n-grams. For mentioned persons we used the pretrained named entity recognition from Spacy in addition to the aforementioned metrics.

To evaluate our automatic algorithm, we randomly chose 50 dialogues and asked an assistant who was not otherwise involved in the project to manually annotate these dialogues. On this, admittedly small, set our automatic method reached high NER precision and recall with 97.8% and 91.9% respectively. The lower recall is mostly caused by typing errors from the crowd workers, so that our algorithm could not detect some of the entities.

named entity resolution	
precision	0.977
recall	0.920
f1-score	0.947

Table 2: Evaluation of our named entity resolution.

3.2.2. Usage of Profile Entities

To show that the crowd-worker really talked about the given profile entities, we computed the overall coverage of named entities. For every dialogue we compared the entities given to the worker in the profile and the detected named entities in the dialogue; counting each match. Averaging over the dialogues, we find that 93.1% of the profile entities are indeed mentioned in a dialogue. (We did not calculate whether *additional* entities may have been mentioned, as we did not want to restrict that from happening.)

3.2.3. Adherence to the Opinion Profiles

Another crucial property is the correct usage of the given opinions. Automatically validating this was not trivial, as it requires co-reference resolution and sentiment analysis for our specific data. We assumed that the effort would be worthwhile, though, as a detailed sentiment analysis would augment the dataset with additional fine-graded information potentially useful for training models with the data (see next section).

To detect an opinion about a named entity, we first had to resolve indirect references (e.g. "*I like it!*" may need to be converted to "*I like Pulp Fiction!*"). We used the coreference resolution annotator from CoreNLP to replace the references with their entity names. First we substituted the original movie titles, person names, countries and genres (as recognised in the NER step) with placeholder words like "*Pulp Fiction*" or "*Peter Pan*" which we confirmed to be recognised by CoreNLP, as it turned out that unusual names or long movie titles are challenging for CoreNLP, especially with typos or lowercased. For our specific case we noticed some problems with the CoreNLP heuristics, presumably because our data is different from its training data. Therefore we manually filtered co-reference chains with first or second person pronouns, as CoreNLP had problems with resolving them correctly and in our case only third person entities are relevant.

To detect the named entity related sentiments, the smallest unit around an entity that can carry a sentiment needs to be identified. An example is given in figure 4. Therefore we used the annotated parse trees from CoreNLP and determined the smallest subordinate clauses within each sentence and all noun phrases with a recursive tree search. In a second step sentence fractions are merged until they contain up to two different nouns. We noticed problems with the annotated parse trees on sentences with grammatical er-

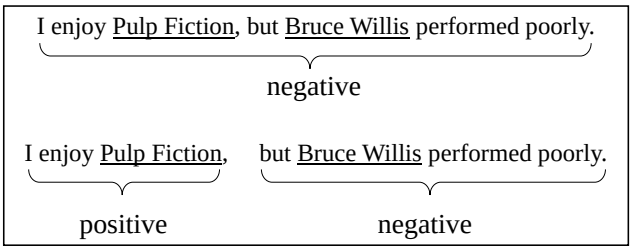

Figure 4: Estimated sentiments from CoreNLP. Named entities are underlined. In the first row Pulp Fiction is declared as negative by mistake.

entity	matches	errors	neutral	accuracy
movie	122	5	36	0.96
person	91	1	35	0.99
other	57	2	28	0.97
sum	270	8	99	0.97

Table 3: Accuracy of crowd-worker adherence to their opinionated profile. Manually evaluated on 50 randomly chosen dialogues.

rors, spelling mistakes or wrong punctuation marks, which led to low recall, as we had to ignore such sentences.
In a final step each subsentence was processed through the sentiment annotator from CoreNLP which provides a discrete probability distribution over five sentiments (*VeryNegative, Negative, Neutral, Positive, VeryPositive*). We compared these labels with the given opinions from the profiles. With that approach 53% of all mentioned entities were labeled as neutral, in 80.1% of the cases, the estimated sentiments conformed with the profile. For a meaningful evaluation of our dataset, the automated approach is not precise enough, so again we evaluated 50 randomly chosen dialogues manually. The results of the manual evaluation are shown in table 3. For most the crowd-worker followed their instructions with a high accuracy of 97%.

To sum up, our analysis showed that the crowd-workers

(i) produced relatively rich and long dialogues (on average, 14 turns),

(ii) talked about the entities they were given as topics, and

(iii) took on the pre-specified opinions about them.

3.2.4. Detailed Sentiment Labels

The validation of our dataset yielded a lot of useful information, which we use to augment our dataset with utterance-level labels regarding entities and sentiments. Later we show in section 5 that these labels can help to improve dialogue quality of our neural network models.

4. Model

To show the contribution of our dataset towards more controlled neural chat generation, we present a baseline model.

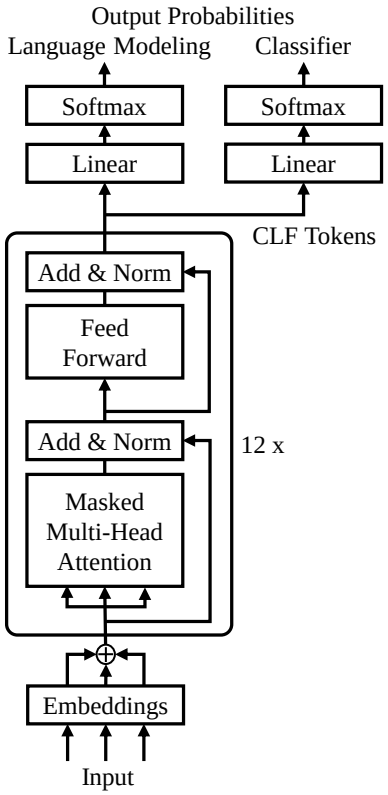

Figure 5: Transformer architecture (Radford et al., 2018) of our baseline model.

The task for the model is to generate the next utterance, given a dialogue with some facts and opinions. It is a generative model trained on two objectives, language modeling with cross-entropy and next-utterance classification. It has the same architecture as the decoder part from Vaswani et al. (2017) with the multi-head attention over the encoder removed, as here we do not have an encoder. This architecture was introduced by Radford et al. (2018) and is commonly known as GPT. It has a stack of 12 identical layers and each layer consists of 2 sub-layers: a masked multi-head self-attention layer and a position-wise, fully connected feed-forward layer with residual connections (He et al., 2016) around them. Like in the original implementation, we used 768 dimensional states and 3072 nodes in the feedforward layers and 12 heads per attention layer. It was used by Wolf et al. (2019) with great success in a similar task on the PERSONA dataset and outperformed state-of-the-art approaches in the Conversational Intelligence Challenge 2.[4] Therefore we used this as our base. Our PyTorch implementation can be found in our repository, mentioned in Section 2., as well. The model is diagrammed in Figure 5.

4.1. Pre-training

We used generative pre-training (Radford et al., 2018) with a selfmade corpus inspired by Zhu et al. (2015). That corpus contains over 1 billion tokens from books across different genres.[5] The model weights Θ are trained given the tokenized corpus $X = \{x_1, x_2, ..., x_n\}$ with minimizing

[4] http://convai.io/

[5] Collection code will be included in the public repository.

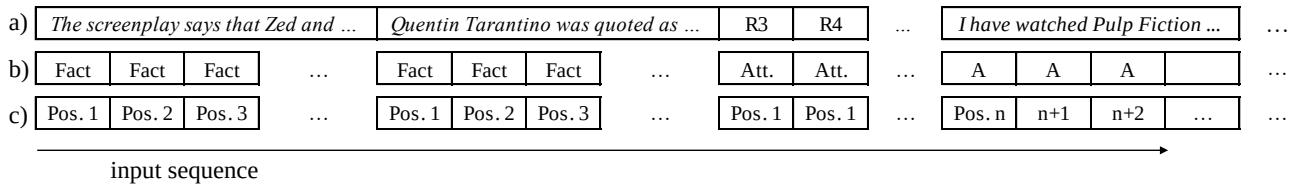

Figure 6: Input representation of our baseline model. The final embedding of each step in a sequence is the sum of all three embeddings: a) The byte paired encoding of the utterances. b) The content encoding that specifies the type of a token. c) The pre-trained positional encodings, shared across the facts.

the negative log likelihood:

$$L_{\text{lm}} = \sum_i -\log P(x_i|x_{i-1}, ..., x_{i-k}; \Theta) \qquad (5)$$

which is a standard language modeling approach. The tokens are split into sequences with length $k = 512$. This unsupervised task allows the model to learn long-term dependencies on coherent text, as well as the token and positional embeddings.

4.2. Training

Before fine-tuning we had to adapt our data so that it fits into the decoder architecture. Similar to Wolf et al. (2019), we decided that for our baseline model, we concatenate facts, attitudes, dialogue history and the next utterance to one input sequence.

In contrast to the pre-training, our setup has a dialogue history, additional facts and attitudes instead of just concatenated sentences. Therefore we need additional input embeddings to represent our more structured data. We used a new set of embeddings which are added to the sum of word tokens and positional embeddings. We used them to differentiate whether a set of tokens (e.g. an utterance) belongs to a specific dialogue partner ('*A*' or '*B*'), a fact or an attitude. The latter are represented with additional embeddings, one for each discrete attitude state. The general concept is shown in Figure 6. Where '*Fact*' and '*Att*' are groups of tokens that differentiate between their targets (e.g. the movie or a specific person). To ensure invariance to the order of the facts and attitudes, the same positional embeddings are used across all additional input, which is also illustrated in Figure 6. Dialogue history, facts and attitudes are concatenated into sequences with a maximum length of 512 tokens. Furthermore we added a classification token at the end of the last utterance, which is ignored by the language modeling loss, but used as input for the classification loss.

After the last hidden layer we multiplied all tokens that did not belong to the last utterance with zeroes to avoid the model learning to predict other tokens than the ones from the last utterance.

To improve generalization, we used delexicalisation for the named entities. That includes movie titles, actors, directors, writer, budget values, age certificates, genres, countries and release years. It is important to note that this step removes the possibility to talk about more than one movie at a time.

We have finetuned the model with a batchsize of 32 for $240,000$ steps on our own dataset, which equals three epochs. After that, both the language modeling loss and the classification loss on our validation set stopped decreasing. A sequence has up to 512 tokens with shorter sequences padded to the maximum sequence length. We used adam optimizier with an initial learning rate $lr = 6.25e-5$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e-8$. We reused most of the parameter from pre-training: General dropout after all layers with $p_{\text{drop}} = 0.1$, weight decay regularization (Loshchilov and Hutter, 2017) with $w = 0.01$ and the new embeddings are initialized with simple weight initialization of $N(0, 0.02)$.

4.2.1. Loss function

In addition to the language modeling loss, described in section 4.1., the model was tasked with identifying the correct next utterance in four candidate sequences $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}, \boldsymbol{x}^{(4)}$. (The rationale for this will be described below.) The wrong sequences were built by concatenating the dialog history with three different utterances from our dataset. Then they are fed, together with a label y, into the model, given a standard classification loss:

$$L_{\text{clf}} = -\log P(y|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}, \boldsymbol{x}^{(4)}; \Theta) \qquad (6)$$

The overall loss to optimize is the sum of both, L_{lm} and L_{clf} with the language modeling loss being reduced by half. Combining both of these losses can help to improve generalization and to accelerate convergence as shown by Radford et al. (2018). In addition, the classification loss can help to refuse at inference time generated sequences which do not fit well as a good answer. This will be explained further in section 4.2.2.

In our first approach these utterances were randomly chosen from different dialogues about the same movie (hereinafter called *random distractors*). In a second step we used the detailed sentiment labels to create wrong utterances that represent a more challenging task. If the correct utterance contains an entity, then false utterances are selected that also contain that entity and have different sentiments, if possible (hereinafter called *rule-based distractors*).

4.2.2. Decoding

We used beam search decoding with a beam size of 4 to generate sequences at inference time, when no ground truth labels are available. To normalize over the length of the sequences we used:

$$lp(Y) = \frac{(5 + |Y|)^\alpha}{(5+1)^\alpha} \qquad (7)$$

	Naturalness	Attentiveness	Consistency	Personality	Knowledgeability
dataset	4.20 (0.96)	4.22 (0.91)	4.36 (0.80)	4.55(0.72)	4.14(0.97)
random distractors	4.01 (0.80)	3.90 (0.93)	4.03 (0.73)	3.86 (0.65)	3.93 (0.72)
rule-based distractors	**4.11** (0.67)	**4.09** (0.71)	**4.05** (0.56)	**4.01** (0.69)	**4.04** (0.63)

Table 4: Human evaluation of our baseline model and our dataset. All 5 categories were evaluated on a likert scale with 5 levels. Standard deviation is shown in brackets.

which is defined in (Wu et al., 2016). With $|Y|$ as the current sequence length and $\alpha = 0.6$ as the length normalization coefficient.

In addition to that, we filtered sequences with multiple identical 3-grams at every beam search step to avoid loops like: '*he performed great and he performed great*' which otherwise is a common occurrence in beam search decoding.

After all possible sequences were found, we combined the generated score with the logits from the classification layer of our model to choose the final sequence. As the classifier loss has learned to distinguish between a correct and two wrong utterances, this gives an additional source for choosing a final beam.

5. Evaluation

In section 3.2. we validated our human/human dataset regarding correct usage of the given profiles. Now we want to evaluate the general dialogue quality for both our dataset and the output of the baseline model. As automated metrics are not very meaningful when used to evaluate the quality of dialogues (Liu et al., 2016), we have performed a human evaluation. The results are shown in table 4. First we explain the used metrics and then evaluate the results regarding our dataset and baseline model.

5.1. Human Evaluation Metrics

For the human evaluation we used Amazon Mechanical Turk again. To evaluate our dataset, we presented pairs of dialogue and one profile to crowd workers to rate. For our baseline model, we asked crowd-workers to chat about a given movie, but did not mention that their chat partner is a bot. We asked the Turker to rate some statements according to their agreement on a Likert scale with five levels from *strongly disagree* to *strongly agree*. The following statements were used:

- Naturalness: The conversation felt natural.
- Attentiveness: It felt like your chat partner was attentive to the conversation.
- Consistency: The conversation was overall consistent.
- Personality: The conversation fits well to the intended character described above.
- Knowledgeability: Your partner replied correct to the asked questions.

Crowd-sourced evaluation may be of low quality, if the crowd-worker are not carefully selected and controlled. Therefore we only accepted crowd-worker with a minimum acceptance-rate of 95% and implemented two fake questions to detect persons that answered randomly. We asked for the subject of the conversation (correct answer is always *movies*), as well as the name of the movie they talked about. If one of the questions was answered incorrectly, we rejected that answer. We evaluated 360 dialogues with 95 different crowd-worker across the three tasks.

5.2. Dataset

The results for our dataset, shown in Table 4, are all above 4 (between *agree* and *strongly agree*), which means that the collected data are judged as natural and consistent dialogues. The high result of 4.55 for personality is consistent with our validation and confirms adherence with the profiles. This and the results from our validation in section 3.2. confirm a senseful dataset with correct labels and natural conversations.

However, the value regarding the knowledgeability is slightly lower as the others. One downside of the movie- and entity restrictions we had while collecting our data is that sometimes the crowd-worker did not know enough about the subjects they were chatting about. If that were true and one asked a random question, their partner was not able to answer this. In general, most of the questions were answered properly though and our model was able to learn this behaviour quite well.

5.3. Baseline Model

We evaluated two variants of our baseline model, one trained with randomly sampled distractors, one with rule-based (sentiment-/entity-sampled) ones (see Section 4.2.1. above). The results are shown in table 4. We also show automated metrics for our model in table 5. The rule-based distractors represent a more difficult classification task at training time and outperformed the random distractor approach in the human evaluation. While both models are nearly equal in naturalness and consistency, rule-based distractors lead to significantly better results in personality and knowledgeability. However, while evaluating both models by our own, we sometimes noticed inconsistencies regarding the opinions. One reason could be that at pre-training the model has learned to condition only on language. As it is much more likely that these utterances were semantically wrong instead of just expressing the wrong sentiment, the model can not learn to distinguish between the different attitudes properly.

model	perplexity	hits@1	hits@3
random distractors	**12.38**	74.22	92.41
rule-based distractors	13.50	**79.33**	**95.51**

Table 5: Automated metric results on our test set. hits@n means n best response selection from 19 additional wrong utterances.

With automated metrics, the approach with random distractors has the better perplexity. That contradicts with the human evaluation, but confirms that automatic metrics do not always correlate with human perception. The hits@n metric though, lines well with the human evaluation. To be comparable, at test time we generated the utterances for both models randomly. The improvement for the rule-based distractors at training time shows, that our additional labels are meaningful and can help to improve the classification task.
The overall results show that it is possible to train an end-to-end chatbot that can not only generate natural answers but also reasonable content, while being consistent to a personality (expressed through opinions to facts) and some external knowledge.

U1: "Hey, how are you?"
S1: "I'm good. How are you?"
U2: "Good!"
S2: "Have you seen Saw?"
U3: "No, what is that movie about?"
S3: "It 's about two strangers who awaken in a room with no recollection of how they get there."
U4: "Do you like Saw?"
S4: "I don't really like the movie."
U5: "I like reading books. How about you?"
S5: "I don't really like the movie."

Figure 7: A dialogue between a human and a model trained on our dataset. The conversation about the movie Saw seems quite natural (U1 to S4), but the model is not able to handle different topics (U5, S5).

6. Conclusion and Future Work

We have presented a new labeled dataset of dialogues, where each dialogue has additional information, namely facts and opinions. This opens a new way to overcome the general problem of inconsistency in end-to-end trained chit-chat models, as we showed with our first baseline model. To be overall consistent, it is important to also be consistently opinionated. With our differentiation of knowledge and opinions, both can be explicitly trained. The baseline model was able to make use of external knowledge in a non-goal driven dialogue, while also representing an opinion and still be natural.
For the future, we are going to explore new model architectures that can handle the additional information in a way different from just concatenating everything as one input sequence. Furthermore, we want to remove the delexicalisation tokens and augment the model with a larger knowledge base, instead of it being restricted to a specific movie. Since our dataset is set in the domain of movies, a model trained on that model is not able to talk about anything outside that domain. It would be interesting to explore if and how it is possible to transfer the property of being opinionated to other, more general dialogue datasets.

U1: "What do you think about Jurassic World?"
S1: "I do not really care for that movie"
U2: "Why?"
S2: "I do not like Irrfan Kahn"
U3: "Who is he?"
S3: "He was signed for the lead role"
U4: "So you do not like the movie?"
S4: "I really do not really care for that one"

Figure 8: A dialogue with the system trained on our dataset, where the system represents a consistent opinion (S1 to S4). However, the grammar of S4 is wrong.

7. Bibliographical References

Ashwini, S. and Choi, J. D. (2014). Targetable named entity recognition in social media. *arXiv preprint arXiv:1408.0782*.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. (1977). GUS, A Frame-Driven Dialog System. *Artificial Intelligence*, 8:155—-173.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2018). Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., and Jurafsky, D. (2016). Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsuper-

vised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Loshchilov, I. and Hutter, F. (2017). Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Ritter, A., Cherry, C., and Dolan, B. (2010). Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics.

Schlangen, D., Diekmann, T., Ilinykh, N., and Zarrieß, S. (2018). slurk–a lightweight interaction server for dialogue experiments and data collection. In *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (AixDial/semdial 2018)*.

Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Stede, M. and Schlangen, D. (2004). Information-seeking chat: Dialogues driven by topic-structure. In Enric Vallduví, editor, *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*, pages 117–124, Barcelona, Spain, July.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Traum, D. R. and Larsson, S. (2003). The information state approach to dialogue management. In Ronnie Smith et al., editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer, Dordrecht, The Netherlands.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Zhou, K., Prabhumoye, S., and Black, A. W. (2018). A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.