MovieChats: Chat like Humans in a Closed Domain

Hui Su[1][*]**, Xiaoyu Shen**[2][*]**, Zhou Xiao**[1]**, Zheng Zhang**[3][†]
Ernie Chang[4]**, Cheng Zhang**[1]**, Cheng Niu**[1]** and Jie Zhou**[1]
[1]Pattern Recognition Center, Wechat AI, Tencent Inc, China
[2]Amazon Alexa AI   [3]Tsinghua University
[4]Department of Language Science and Technology, Saarland University
`aaronsu@tencent.com,gyouu@amazon.com`

Abstract

Being able to perform in-depth chat with humans in a closed domain is a precondition before an open-domain chatbot can ever be claimed. In this work, we take a close look at the movie domain and present a large-scale high-quality corpus with fine-grained annotations in hope of pushing the limit of movie-domain chatbots. We propose a unified, readily scalable neural approach which reconciles all subtasks like intent prediction and knowledge retrieval. The model is first pretrained on the huge general-domain data, then finetuned on our corpus. We show this simple neural approach trained on high-quality data is able to outperform commercial systems replying on complex rules. On both the static and interactive tests, we find responses generated by our system exhibits remarkably good engagement and sensibleness close to human-written ones. We further analyze the limits of our work and point out potential directions for future work [1].

1 Introduction

Being able to converse like humans in a *closed domain* is a precondition before an intelligent open-domain chatbot, which further requires transiting among various domains, can be designed (Gao et al., 2019; Su et al., 2020). Nonetheless, even if constrained in a specific domain, current chatbots are still far from satisfactory. Unlike task-oriented systems that can be relatively well-resolved with handcrafted templates, human conversations feature a complex mixture of QA, chitchat, recommendation, etc. without pre-specified goals or conversational patterns (Dodge et al., 2016; Akasaki and Kaji, 2017; Shen et al., 2018). Selecting proper

domain knowledge to support response generation at all the different situations is challenging (Milward and Beveridge, 2003; Lian et al., 2019; Shen et al., 2019). In this work, we direct our focus to the movie domain and present a *large-scale, crowd-sourced Chinese dataset with fine-grained annotations* in hope of boosting the study towards a human-like closed-domain chatbot.

A variety of dialogue datasets with grounded domain knowledge have already been proposed. However, they are collected either through (1) online forum crawling (Dodge et al., 2016; Ghazvininejad et al., 2018; Liu et al., 2018; Zhou et al., 2018a; Qin et al., 2019), which are noisy, multi-party, mostly contain only single-exchange QA, or (2) crowd-sourced (Zhu et al., 2017; Zhou et al., 2018b; Moon et al., 2019; Wu et al., 2019), which are small-scale and often created in an overconstrained setting like teacher-student (Moghe et al., 2018). Even for datasets crowd-sourced in unconstrained scenarios, suggestive domain knowledge is provided for humans before an utterance is provided. This would inevitably prompt humans to utilize these knowledge deliberately, yielding unnatural conversations simply connecting the knowledge (Dinan et al., 2019; Zhou et al., 2020). We show examples from other datasets in Appendix Table 10. In comparison, our dataset has the following advantages:

1. **Natural**: Crowdworkers chat in a free environment without further constraint or prompt in order to mimic the human daily conversations to the largest extent.

2. **Large-scale**: It covers 270k human dialogues with over 3M utterances, which is *at least one order of magnitude larger* than all the other crowd-sourced datasets.

3. **Annotated**: Utterances are labeled with entity information and dialogue acts classified into

15 fine-grained aspects, based on which linked into different types of knowledge.

Different from previous crowd-sourced works, our annotation process is conducted posteriori so that it will not interfere with human conversations, e.g., prompt them to overuse suggested knowledge.

Built upon our dataset, we propose a *simple unified language model approach* to push the limits of movie-domain chatbots. The model is first pretrained on 2.2B words collected from various general-domain conversational resources, then fine-tuned on the movie dataset with additional knowledge and dialogue acts incorporated. We pool all components like intent prediction and knowledge retrieval into a sequence prediction task and solve them with a unified language model architecture. It avoids designing complex systems for individual components separately and all subtasks can be easily trained simultaneously (Hosseini-Asl et al., 2020; Peng et al., 2020). We show our simple unified approach outperforms strong baselines for each separate subtask. Knowledge retrieval, dialogue acts prediction and general-domain pretrain benefit from each other and altogether bring improvement to the generation quality. In the online interactive test, our best model succeeds at chatting with humans for 11.4 turns without being detected to be a machine, outperforming even commercial chatbots Mitsuku [2] and Microsoft XiaoIce [3] which further rely on complex rules. By analyzing the limitations of our model, we find it especially has difficulty at dealing with in-depth discussions over long turns. Future research can consider employing larger knowledge base or explicit state tracking.

In summary, our main contributions are (1) presenting a high-quality, large-scale Chinese conversational corpus with fine-grained annotations in the movie domain to benefit future study, (2) showing that a simple unified neural model trained on the high-quality dataset can approach human performance and even outperform commercial systems replying on complex rules, and (3) studying the shortcomings of current techniques, providing suggestive directions for future research.

2 Dataset Construction

The dataset construction consist of (1) crowd-sourcing the dialogues, (2) annotating dialog acts and entities and (3) linking utterances into

grounded knowledge. We explain these three steps in order and present the dataset statistics in the end.

Dialogue Crowd-sourcing We obtain the dialogue dataset through a two-phase Wizard-of-Oz-style collection (Kelley, 1984; Dahlbäck et al., 1993). In the first phase, we run small-scale pilot studies and examine the quality of collected conversations. Based on the examination, we created tutorials and qualification tests. They are used to train and qualify crowd-workers for the second phase. During this second phase, we consistently monitor the collected dialogue datasets and perform periodic quality check on samples from every individual work pairs. If more than 5% from one pair are considered invalid, their collections will be removed. Before a conversation started, two workers are paired and a movie is chosen agreed by both [4]. We constrain at least one of them to have watched the movie to make sure the conversation is contentful [5]. The annotators are especially instructed to (1) behave naturally as in daily life, (2) avoid dirty words and (3) talk differently in each conversation. Duplicate conversations will be removed if more than 70% of their contents are overlapped. To encourage diverse movies, we further set an upper limit to forbid one movie from being talked about for more than 100 times.

The whole collecting process lasts two months. In the end, 245 participants are involved with 66,424 movies being talked about in total.

Dialogue Act and Entity Annotation Following prior work, we base our annotation schema on the ISO 24617-2 standard (Bunt et al., 2010, 2012). Table 1 shows our annotation schema, counts, descriptions, and brief examples. The dialogue acts (DAs) are organized in a hierarchical structure. The first layer makes distinctions on three concepts: objective facts, recommendations and subjective feelings. Each concept can either be either requested or informed during the conversation. We further define an "*Other*" class to include actions that do not belong to any of the three concepts, like some general non-contentful greetings or echos. The second layer includes 15 finer-grained aspects covering most popular topics being discussed. Every first-layer DA (except *Other*) will be further group it into one of these 15 aspects, e.g., the de-

[4]We fix one movie beforehand as the main topic, but they are allowed switch to other movies as the conversation goes.

[5]The knowledge of the movie is presented to the worker who has watched it to refresh his memory about movie details.

我刚看完《海上钢琴师》，感觉心灵很震撼
I just finished watching "*The Legend of 1900*" and feel shocked

DA: Inform_fact, Inform_feeling
Aspects: Name
Mov_Tracker: *The Legend of 1900*

Name: The Legend of 1900 | **Comment:** *One of the favorite movies that shocked my soul.*

DA: Request_fact
Aspects: Type
Mov_Tracker: *The Legend of 1900*

真的吗，这是部什么电影
Really, what kind of movie is this

一部意大利拍的英语电影，讲一位天才钢琴师一辈子都住轮船上
An English film made in Italy, where a talented pianist lived on the ship all his life

DA: Inform_fact
Aspects: Region, Language, Plot
Mov_Tracker: *The Legend of 1900*

Region: Italy | **Language:** English | **Plot:** "a talented pianist lived on the ship all his life"

DA: Inform_feeling, Request_fact
Aspect: Plot
Mov_Tracker: *The Legend of 1900*

这么离奇的故事，为什么他不离开那艘船呢
Such a bizarre story, why didn't he leave that boat?

Plot: "why didn't he leave that boat"

他把这里当成了他的精神家园，所以船要被炸掉他都不肯离开
He regarded this as his spiritual home, so when the last ship was about to be blown up, he refused to leave

DA: Inform_feeling, Inform_fact
Aspects: Plot
Mov_Tracker: *The Legend of 1900*

Plot: "when the last ship was about to be blown up, he refused to leave"
Comment: *What he cannot leave is not the boat, but the spiritual home that nurtured him.*

DA: Inform_feeling
Aspects: Name
Mov_Tracker: *Forrest Gump*

就像阿甘正传一样。现代社会很少有人能和世俗割裂去追求自己的精神世界
Just like in *Forrest Gump*. Few people in modern society can separate themselves from the world to pursue their own spiritual world

Name: Forrest Gump | **Comment:** *The negative perspective of 1900 may represent the director's reflection on modern civilization.*

Figure 1: An example of our dataset. The annotations include dialogue act, aspects and movie tracker and grounded knowledge.

tailed DA of the first example in Table 1 will be request_fact_director. If one utterance contains multiple dialogue acts, we order the dialogue acts based on their turn of appearance in the utterance. As for the named entity recognition, we labeled 5 kinds of entities: movie names, director, actor, type and role (first 5 aspects).

To speed up the annotation process, we first define a set of handcrafted regular expressions, which covers most frequent patterns at each class, to train a DA and NER classifier . The annotators are instructed to post-correct the auto-labeled dialogues instead of doing everything from scratch. The classifiers are trained with online learning (Sahoo et al., 2018) to keep improving the accuracy and lower down the frequency of post-correction in consequence. As we observe, this semi-automated way significantly speeds up the labeling process. All the dataset is finished labeling within three weeks with 188 annotators involved.

Knowledge Linkage We extract fact knowledge from the structured table in Douban Movie [6] , a popular Chinese platform for movies. The knowledge is organized in the form of key-value pairs, where the key corresponds to the 15 aspects defined by us. Some aspects, like lines or music, are not directly available from the structured table. We extract these missing information from other sources and combine it into our knowledge base. For utterances labeld as *inform/request_fact*, we link them to the key-value pairs from the same aspect. Apart from the objective knowledge, we also crawl movie comments from Douban Movie to support the generation of responses expressing subjective feelings. These comments can be a good supplementary to provide knowledge that can be hardly organized in the structured form (Moghe et al., 2018). For utterances labeled as *inform/request_feeling*, we compare them with Douban comments from the same movie and compute the similarity score based on weighted average of edit distance, Jaccard distance, tf-idf, sentence vector cosine similarity, common words and entities. Each utterance is linked to the most similar comment with a threshold cutoff. In the end, 51.7% of the utterances about feelings have grounded comments. For utterances about recommendations, we simply ground them to the men-

[6]https://movie.douban.com

6607

Dialogue Act	Count(%)	Linked	Description	Example
Request_fact	8.62	Fact	Request facts.	Who directed this movie?
Request_recommend	4.91	None	Ask recommendations.	Which other movies do you recommend?
Request_feeling	4.98	Comment	Request feelings.	How do you like its theme music?
Inform_fact	24.85	Fact	Inform facts.	Wong Kar-Wai directed this movie.
Inform_recommend	4.56	Movie	Give recommendations.	I can also recommend *Titanic*!
Inform_feeling	28.95	Comment	Convey feelings	Its music reminds me of my childhood!
Other	23.10	None	Greetings, echos, etc.	hahaha.

Aspects: name, director, actor, type, role, region, time, plot, line, award, gross, rating, website, music, others

Table 1: Counts, type of linked knowledge, descriptions and examples of the dialogue acts. *Examples are shown only in translated English for space limit*. Each dialogue act (except *other*) is paired with one of the 15 aspects. Details descriptions and statistics of the aspects are in Appendix Table 11.

tioned movie entities [7], and no grounded knowledge is linked for utterances labeled as *Other*. An example of our annotation is presented in Table 1.

	Train	Valid	Test/s	Test/u
Dialogues	221,526	12,307	6154	6154
Utterances	2,702,618	153,591	79,978	74,463
Turns/D	12.2	12.5	13.0	12.1
Tokens/U	14.9	12.6	13.5	15.9
Movies	59,449	3,146	1454	1454

Table 2: Number of dialogues, utterances, turns per dialogues, tokens per utterance and movies in the train, valid, test (seen movies) and test (unseen movies) dataset.

Dataset Statistics The final dialogue dataset we collect consists of 246,141 dialogues with 3,010,650 turns, which we divide into 221,526 for train, 12,307 for validation, and 12,308 for test. The test set is split into test/seen for movies seen in the training set, and Test/unseen for unseen movies, each covers conversations about 1454 movies. The training set covers 59,449 movies with 12.2 turn per dialogue in average. Each turn contains 14.9 tokens. Data statistics can be found in Table 2.

3 Model Architecture

Language models have demonstrated impressive performance as a universal learner across NLP tasks (Shen et al., 2017; Peters et al., 2018; Radford et al., 2019; Brown et al., 2020). Inspired by this, our dialogue generation model is implemented as a Transformer-based language model like GPT2 (Radford et al., 2019; Zhang et al., 2019). It contains a pipeline process of movie tracker, intent prediction, knowledge retrieval and text gener-

---

[7]We only consider recommending *movies* as for the DA about recommendation. Recommending other aspects require assembling recommendation systems of different domains, which is beyond the scope of this paper.

Context	[context] *dialogue context*
Fact	[fact] *key-value pair(s)*
Comment	[comment] *movie comment(s)*
Recommend	[recommend] *movie name(s)*
Track	[tracker] *[inherit] or a new movie name*
Intent	[intent] *DA sequence*
Retrieve	[retrieve] *knowledge*
Response	[response] *response*

Table 3: A schematic representation of the different components. Upper parts are the conditions and lower parts are the targets. The condition and target are concatenated into a single sequence then fed into the language model.

ation. Unlike in traditional task-oriented systems where subtasks are decomposed separately, we opt for *a simple and unified approach by casting all subtasks into sequence prediction*. A special token is injected in the beginning to indicate which subtask to perform (Hosseini-Asl et al., 2020; Peng et al., 2020). Table 3 shows the schema representation for different components. The condition and the target are concatenated into a single sequence and then fed into the language model to train. For example, the task of predicting the intent given the dialogue context will be transformed into "[context] *dialogue context* [intent] *DA sequence*", where the *DA sequence* will be predicted conditioned on "[context] *dialogue context* [intent]".

General-domain Pretrain The model is first pretrained on a mixed general-domain conversational corpus crawled from various sources like

douban [8], tieba [9], zhihu [10] and weibo [11]. The pretrained corpus covers 468M conversations with 2.2B words. Each training instance is processed in the form of "[context] *dialogue context* [response] *response*" where *response* will be predicted given "[context] *dialogue context* [response]". The objective is a mixture of maximum likelihood and unlikelihood training (He and Glass, 2019; Li et al., 2019), which we find help reduce repeated and incoherent generations as observed in Adiwardana et al. (2020). The unlikelihood training minimizes the likelihood of 1) randomly sampled responses from the corpus and 2) repeated bigrams from the previous generated tokens.

Movie Tracker   The movie tracker is like the belief state tracker in task-oriented systems (Henderson et al., 2013). It is used to track which movie will be talked about in the next utterance. Conditioned on the dialogue context, we feed a special token [tracker] to the decoder. The decoder should decide whether to inherit the movie talked about in the last utterance or switch to a new movie. The target would be "[tracker] [*inherit*]" or "[tracker] *new movie name*" if transiting to a new movie. The initial state is an empty indicator [*None*].

Intent Prediction   The intent prediction is also cast as a sequence prediction task. Compared with the traditional way of multi-label classification, casting it as sequence prediction is better at addressing the coexistence of multiple DAs and capturing the sequential dependencies among the hierarchy (Raffel et al., 2019; Vedula et al., 2020). For example, to predict the DAs of the 4th utterance in Figure 1, the sequence fed to the language model will be "[context] *dialogue context* [intent] *inform, feeling, plot, request, fact, plot*". By this means, before predicting a DA, the model can condition on both the dialogue context and its previous DAs to improve the accuracy.

Knowledge Retrieval   The knowledge retrieval component is similar to the classical DSSM model (Huang et al., 2013). We replace the MLP with our language model encoder to get the embedding for knowledge. Note that we only select knowledge from the current movie, which can be obtained from the movie tracker, so it is possible to

Parameters	Layers	Dimension	Heads
117M	12	768	12
345M	24	1024	16
762M	36	1280	20

Table 4:   Architecture hyperparameters for the 3 model sizes.

run exact softmax over all knowledge candidates and maximize the likelihood of the ground truth. We condition on both the dialogue context and the intent. The sequence fed to the language model would be "[context] *dialogue context* [intent] *DA sequence* [retrieve] *knowledge*" where *knowledge* will be predicted. When an utterance is not linked with any knowledge, it will predict a *None* token.

Text Generation   The text generation is conditioned on the dialogue context, intent and the retrieved knowledge. All the conditions are concatenated into a long sequence. The knowledge can be fact, comment, movie names of mixture of them according to the DA of the utterance. For example, if grounded on comment knowledge, "[context] *dialogue context* [intent] *DA sequence* [comment] *comment* [response]" will be fed to the language model to generate the response. To make it consistent with the pretrained general-domain dialogue, the position embedding of the decoded response will skip the concatenated intent and knowledge and directly follow the dialogue context. We find this beneficial when combined with pretrained models. The objective also follows the pretrained model mixing maximum lilkelihood and unlikelihood training.

4   Experiment Setting

We tokenize the Text in the unit of Chinese characters and keep all unique non-Chinese unique tokens appearing for more than 5 times. The whole vocabulary contains 13,317 words. We train our model on 24 Nvidia V100 GPUs (32GB) with three different model sizes as shown in Table 4. The batch size is fixed as 64 per GPU. The context length is truncated to be 300 words. We optimize models with the Adam gradient descent and a drop out rate of 0.1. For the 117/345M model, the learning rate is set as $1.5e - 4$ with a linear warm up of 3600 steps. For 762M model, the learning rate is $1.25e - 4$ with a 2400-step linear warm up. Models are trained for 150k steps on the pretraining corpus and we observe no overfitting for all. For decoding, we employ Nucleus Sampling with $p = 0.9$ (Holtzman et al., 2020). All experiments are done with

---

[8]https://www.douban.com/
[9]https://tieba.baidu.com/
[10]https://www.zhihu.com/
[11]https://www.weibo.com/

Model	seen						unseen					
	Perplexity	BLEU-1/2/3			Distinct-1/2		Perplexity	BLEU-1/2/3			Distinct-1/2	
117M	26.42	24.71	2.07	0.44	10.6	25.4	31.79	19.12	1.95	0.34	6.7	26.8
+DA	27.33	25.12	1.99	0.62	9.1	24.7	32.22	16.63	2.11	0.41	7.5	28.1
+Aspect	24.29	25.06	1.85	0.86	9.5	28.2	30.36	17.57	2.36	0.39	9.6	22.4
+Knowledge	20.25	29.21	2.36	1.34	8.7	26.3	28.61	24.48	2.71	0.57	10.7	27.5
+Pretrain	18.06	33.46	2.68	1.63	10.6	28.5	26.59	28.12	2.62	0.95	11.3	32.0
345M	24.74	28.06	2.17	1.40	10.0	30.7	29.61	18.48	2.36	0.49	12.6	26.6
+DA	24.52	22.28	6.61	3.50	9.7	22.6	28.26	16.67	2.95	0.89	12.8	24.3
+Aspect	23.61	25.51	**6.70**	2.08	12.5	27.9	28.55	16.97	3.55	0.93	13.2	20.7
+Knowledge	18.64	27.01	4.27	2.89	13.4	32.2	24.61	21.13	2.58	1.08	13.8	32.8
+Pretrain	15.76	33.23	6.73	3.12	11.6	29.9	20.27	27.03	4.35	1.56	13.5	33.1
762M	19.53	28.69	5.45	2.44	11.4	**37.3**	25.83	22.25	3.90	1.04	12.7	31.3
+DA	19.33	31.25	4.27	2.89	11.1	35.5	25.67	24.95	3.59	2.75	14.4	26.9
+Aspect	18.94	31.05	6.01	2.67	7.8	23.8	24.75	23.84	3.61	2.09	**16.8**	31.7
+Knowledge	16.45	33.55	5.53	2.71	12.6	31.9	20.30	25.18	4.39	2.17	13.8	30.9
+Pretrain	**13.17**	**38.33**	5.88	**4.02**	**15.1**	37.2	**17.49**	**29.02**	**5.79**	**2.65**	15.6	**36.8**

Table 5: The perplexity, BLEU score and distinct uni/bi-grams of our model with different sizes. Results are reported for seen/unseen movies respectively. Each line adds one more condition on top of the above one.

PyTorch (Paszke et al., 2019).

5 Results and Analysis

Automatic Evaluation In Table 5, we report the perplexity, BLEU scores and distinct uni/bigrams for three model sizes. To investigate the effects of incorporating annotations and pretraining, we start from a basic model which trains from scratch on our movie corpus. At each time, we add one more condition to see its influence. The results show a clear tendency of gradual improvement as more conditions are added to the training. Adding knowledge especially boosts the performance, which is understandable considering movie-domain chats usually contain many movie-specific rare names. Without knowledge grounding, it can hardly predict the correct tokens. Pretraining on general-domain conversations can improve both the overlap with ground truth. The distinct uni/bigrams also consistently increase, implying the model can learn useful patterns in the pretrained corpus to enrich its generations in the movie domain. In unseen testset, the performance generally drops for all, especially for models without knowledge grounding as they have to make up facts and comments for totally unseen movies in the training set.

Table 6 measures the accuracy of predicting dialogue act (DA), aspect and movie tracker of our model. Our models are all pretrained with general-domain corpus beforehand. Apart from being trained only to predict the individual tasks, we include the results where all subtasks are co-trained end-to-end in the last line. We compare our models with the Chinese RoBERTa (Liu et al.,

Model	DA1	DA2	Aspect	Movie
RoBERTa	75.8	64.0	44.3	90.7
117M	76.7	66.3	43.8	88.9
345M	82.5	71.5	46.1	90.3
762M	82.9	73.4	51.4	**92.1**
762M (co-train)	**84.5**	**76.5**	**54.3**	91.7

Table 6: Accuracy of dialogue act, aspect and movie tracker.

2019), which has achieved SOTA performance on Chinese NLU tasks. We use its optimized version for Chinese which is pretrained by masking full Chinese words (Cui et al., 2019) [12]. RoBERTa does the movie tracking by performing softmax over the recognized named entities. For dialogue acts, we deal with DA1 (inform/request/other) and DA2(fact/feeling/recommend) separately. As can be observed, our simple unified approach for all subtasks does not come at the expense of the accuracy. When all subtasks are co-trained, the performance is further boosted, indicating the multitask training paradigm can benefit individual tasks.

Model	Fact Hit@1	Comment Hit@1	Recommend Hit@1/5	
Random	0.054	0.011	0.000	0.000
BOW	0.135	0.048	0.011	0.014
Bert	0.406	0.067	0.048	0.075
117M	0.402	0.053	0.045	0.071
345M	0.455	0.071	0.047	0.072
762M	0.518	0.120	0.053	0.086
+ DA	**0.543**	**0.158**	**0.066**	**0.104**

Table 7: Hit rates of knowledge retrieval.

Table 7 measures the performance of retrieving fact knowledge, movie comments and recommen-

---

[12]https://huggingface.co/hfl/chinese-RoBERTa-wwm-ext

dation respectively. We report the hit@1 and hit@5 scores for them (Zhang et al., 2018). We compare our model with a random baseline, bag-of-word (BOW) and the Bert (Devlin et al., 2019) model (we pass sentences through Bert and derive a fixed-sized vector by averaging the outputs from the second-to-last layer (May et al., 2019)). The BOW and Bert model are finetuned with our knowledge linkage annotations. We find that our unified model again outperforms all baseline approaches. Adding the DA as a condition further helps. Fact retrieval has the highest hit rate as it is well structured and easy to match. Recommendation, on the other hand, is very hard to predict. As an accurate recommendation system is clearly beyond the scope of this paper, it is understandable that our simple way fails to provide satisfying recommendations.

Metric	Agreement (%)	Fleiss'k score
Sensibleness	71.5 ± 3.4	0.473
Engagement	64.2 ± 2.3	0.438
Informativeness	79.2 ± 3.1	0.576

Table 8: Crowd worker agreement scores.

Human Evaluation Automatically evaluating dialogue systems are known to be extremely hard (Liu et al., 2016; Su et al., 2018). We further conduct a set of static and interactive human evaluations. We focus on evaluate the machine-generated response from four perspectives. Apart from the oft-used metrics (1) **Sensibleness (Sens)** and (2) **Engagement (Enga)** for open-domain chatbots, we further evaluate on (3) **Factuality (Fact)** and (4) **Informativeness (Info)** to see if models can actively provide informative responses based on movie facts. Details are in Appendix B. As evaluating factuality requires specific movie knowledge, this metric is only evaluated by the same person who produced the dialogue. The other metrics are evaluated by 3 workers each. Table 8 shows the agreement scores. The agreement is reasonable considering the evaluations are subjective. The results are the majority votes of the binary scores.

In the static evaluation, we sample 300 responses for each model from the test set (mixing seen and unseen). The responses can come from any turn in a conversation. We show the results in Figure 2. Our largest model with 762M is clearly preferred by human evaluators on almost all metrics and approaches human performance. By training a larger model and increasing the training size, the gap

might be further closed.

In the interactive evaluation, humans can chat with any topic but restricted in the movie domain. We conduct an online Turing test where one side is always a human participant not aware whom he is talking with. The other side could be either Mitsuku, XiaoIce [13], our model (762M with pretraining) or a real human. Mitsuku interacts in English, so we hire only English native speakers for the experiment. We collect 100 conversations for all models. Humans can stop interacting once they (1) find the other side is a machine or (2) reaches the maximum turn of 20. Responses from all models are later passed to the third party to judge the scores. The results are shown on the right of Figure 2. Our model outperforms Mitsuku and XiaoIce by a large margin. As Mitsuku and XiaoIce are designed to be open-domain chatbots, restricting to be on the movie domain will give our model some natural advantage. We can also notice that Mitsuku and XiaoIce *almost never produce fake facts*. The cost is the extremely low ratio of informative responses since they tend to behave over-safely and will only answer it when they are 100% sure. Our model is closer to humans in that sense. It will converse actively at some risk of containing fact errors.
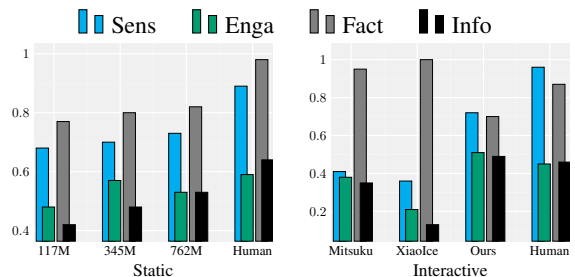

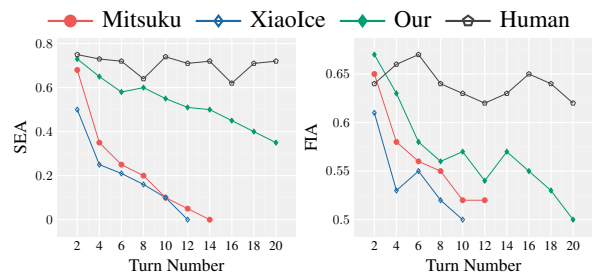Figure 2: Results of static/interactive human evaluation.

Figure 3: Change of SEA and FIA as the turn proceeds.

---

[13] We use its chat service through Weibo. It will sometimes generate responses containing keywords like "XiaoIce". We manually replace it to prevent disclosing its identity.

1	*User Input*	变形金刚看了没，超酷的呢(Have you watched Transformers? So cool)	
	MovieChats	[Others] [None] 当然看过啦(Sure)	
	User Input	我最喜欢大黄蜂了(I like Bumblebee so much)	
	MovieChats	[inform_feeling_role] [变形金刚里大黄蜂最可爱] 是啊大黄蜂真的太太太可爱了叭(Yes, Bumblebee is sooo cute)	
	User Input	可不嘛偷偷把车停在好风景的地方帮山姆撩妹(Exactly, stopped the car in a nice view and helped Sam hit on girls)	
	MovieChats	[inform_feeling_plot] [None] 哈哈，霸天虎当时也在(Haha, Decepticons were there, too.)[factuality=0]	
2	*User Input*	三傻大闹宝莱坞把我看哭了(I was moved into tears by 3 idiots)	
	MovieChats	[inform_feeling_others] [感人的影片，看了很多遍] 我也是，看了一百遍(Me too, have watched it one hundred times)	
	User Input	男主是真的很聪明(The leading actor is really smart.)	
	MovieChats	[inform_feeling_role] [...主角敢于对传统教育说不...] 很喜欢男主，传统教育糊弄不了他(I like him very much. He is not fooled by the traditional educational system.)	
	User Input	所以学校教育模式真的很重要(So the educational pattern at school is very important)	
	MovieChats	[inform_feeling_others] [None] 社会要进步，教育就要不断革新(Education must keep going for social progress)	
	User Input	男主是谁演的(Who plays the actor in the movie)	
	MovieChats	[inform_fact_actor] [actor: 阿米尔·汗...] 阿米尔·汗，非常出名的(Aamir Khan, very famous)	
	User Input	他还演过什么电影(Which other movie did he play)	
	MovieChats	[inform_fact_name] [name: 三傻大闹宝莱坞] 三傻大闹宝莱坞，也很好看(3 idiots, also a nice movie)[sensibleness=0]	
3	*User Input*	我最近在重温哈利波特系列(I'm reviewing Happy Potter series recently.)	
	MovieChats	[others] [None] 你喜欢看吗(Do you like it)	
	MovieChats	[inform_feeling_others] [第一部好看啊] 第一部是我最喜欢的(The first part is my favorite)	
	MovieChats	[inform_feeling_role] [小哈利颜值太高了] 喜欢小哈利，那时候颜值太高了(I like little Harry, he was so cute that time)	
	MovieChats	[request_fact_role] [role: 哈利波特...] 哈利波特是谁演的呀(Who played the role of Harry Potter)	

Table 9: Examples of interactions with our model. First two are failing cases from the interactive evaluation. Last one shows how the response can be controlled by feeding different DA and aspects.

Distance from Human Performance In the interactive evaluation, compared with human performance, our model loses a bit on sensibleness and factuality but wins on the other two. To investigate where our model fails, figure 3 visualizes the change of **SSA** (Sensibleness-Engagement average) and **FIA** (Factuality-Informativeness average) when the conversational turn proceeds. A good chatbot should balance well these skills (Adiwardana et al., 2020). SEA can reflect how it behaves as a general chatbot while FIA can better test its capability at incorporating domain knowledge. We can see a clear trend of decrease for all models. As for human performance, however, the score is quite consistent across turn rounds, implying a large improvement space for current models to **deal with multi-turn context**.

In figure 4, we further show the "dying distribution" of our model, namely, in which DA our model fails to pass the Turing test and thereby "dies". Unsurprisingly, we can see the system fails mostly when informing facts or feelings. Only a small portion are from non-grounded chitchats (*other*). This suggests the most crucial bottleneck lies in the **interaction with movie-specific knowledge and seamlessly incorporating it into the response generation**. We show some snippets of interactions with our model in Table 9. The first two
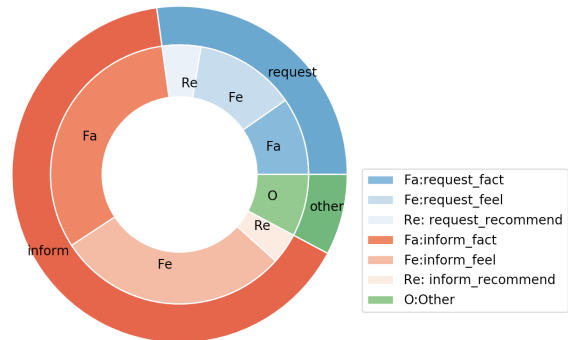

Figure 4: Dying distribution of DAs in interactive evaluation.

are failing cases labeled by humans as not factual and sensible. We can see the model struggles at replying to too specific facts. This is understandable since our knowledge base only provide short introductions and cannot cover all what happened in the movie. The second case shows its shortcoming at handing long-range consistency. It still recommends the current movie when the user asks about "which *other movie*". Employing larger knowledge bases and explicitly tracking the states by a checklist (Kiddon et al., 2016) might potentially alleviate both issue. We also provide examples for controllable generations where the DA and aspect are manually assigned. As observed, the model shows decent performance at fitting both the dialogue con-

text and specified conditions. This can be helpful when finer-grained control is needed.

6 Conclusion

We present MovieChats: a movie-domain chatbot built upon a large-scale, high-quality conversational corpus with fine-grained annotations. The model can be trained end-to-end with a simple unified language model architecture. We show that our model, powered by well-defined knowledge grounding, is able to approach human performance in some perspective, though still lagged behind when it comes to dealing with detailed knowledge or long-turn consistency.

Acknowledge

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Satoshi Akasaki and Nobuhiro Kaji. 2017. Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1308–1319.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an iso standard for dialogue act annotation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*, pages 193–200.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2016. Evaluating prerequisite qualities for learning end-to-end dialog systems. *ICLR*.

Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tianxing He and James Glass. 2019. Negative training for neural dialogue response generation. *arXiv preprint arXiv:1903.02134*.

Matthew Henderson, Blaise Thomson, and Steve Young. 2013. Deep neural network approach for the dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 467–471, Metz, France. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *ICLR*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM*

international conference on Information & Knowledge Management, pages 2333–2338.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2019. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. *arXiv preprint arXiv:1911.03860*.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5081–5087. AAAI Press.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628.

David Milward and Martin Beveridge. 2003. Ontology-based dialogue systems. In *Proc. 3rd Workshop on Knowledge and reasoning in practical dialogue systems (IJCAI03)*, pages 9–18.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model. *arXiv preprint arXiv:2005.05298*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Doyen Sahoo, Quang Pham, Jing Lu, and Steven CH Hoi. 2018. Online deep learning: learning deep neural networks on the fly. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2660–2666.

Xiaoyu Shen, Youssef Oualil, Clayton Greenberg, Mittul Singh, and Dietrich Klakow. 2017. Estimation of gap between current language models and human performance. *Proc. Interspeech 2017*, pages 553–557.

Xiaoyu Shen, Hui Su, Wenjie Li, and Dietrich Klakow. 2018. Nexus network: Connecting the preceding and the following in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4316–4327.

Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. 2019. Select and attend: Towards controllable content selection in text generation. 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)*, pages 579–590.

Hui Su, Xiaoyu Shen, Pengwei Hu, Wenjie Li, and Yun Chen. 2018. Dialogue generation with gan. In *AAAI*.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7087–7097, Online. Association for Computational Linguistics.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2020. Open intent extraction from natural language interactions. In *Proceedings of The Web Conference 2020*, pages 2009–2020.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.

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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4623–4629. AAAI Press.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. *ACL*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713.

Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *arXiv preprint arXiv:1709.04264*.

Online Forum Liu et al. (2018)	A: 求推荐励志电影，谢谢大家 (Any inspirational movies? Thanks everyone). B: 《当幸福来敲门》 (*The Pursuit of Happyness*).
Crowd-sourced (**Constrained**) Zhou et al. (2018b)	A: Hey have you seen the inception? B: No, I have not but have heard of it. What is it about ? A: It's about extractors that perform experiments using...
Crowd-sourced (**unconstrained**) Zhou et al. (2020)	A: 知道重庆森林这个电影吗 (Do you know the movie *Chungking Express*)? B: 知道啊，是王家卫导演的 (Yes, it's directed by Wong Kar-Wai). A: 具体是哪年上映的你还记得吗？ (Remember which year it was on)? B: 记得，是在1994年07月14日 (Yes, July 14th, 1994)...
Ours	A: 威尔史密斯演技真的很棒 (Will Smith's acting skill is really good). B: 他的当幸福来敲门太经典了 (His The Pursuit of Happyness is a classic). A: 一直都挂在电影排行榜靠前的位置 (That' always among top ranked movies). B: 嗯嗯，这部电影真的很励志啊 (Yes, it's really motivational). A: 威尔史密斯也演出了很惨的感觉了 (Will Smith plays like he is a real tragedy). B: 演技特别好 (Yes, he acts pretty well).

Table 10: Comparison between our dataset and others.

A Dataset Collection

Table 10 shows examples comparing our dataset and the others. As observed, forum conversations are mostly single-turn QA or comments. Current crowd-sourced datasets are either collected on constrained scenarios (the scenario in (Zhou et al., 2018b) fixed the roles in a conversation as one introducer and one listener), or unconstrained but prompting people to deliberately connect knowledge. Our dataset simulates real-life conversations to the largest extent.

We classify the utterances into one of 15 aspects. The definitions, counts, and examples of them are shown in Table 11. When annotating the corpus, tutorials and examples are provided to the annotators, we show some examples of in the following tables. All the examples are provided only in Chinese as that is the native language among annotators.

Aspects	Count(%)	Description	Example
电影名(name)	17.41	Name of a movie.	Its name is so funny
导演(director)	2.44	Director of a movie.	Who played Jack in the movie?
演员(actor)	9.24	Actor of a movie.	I wonder who picked him as the leading actor.
类型(type)	7.12	Type of a movie.	I don't really like horror movies.
角色(role)	10.01	Role of a movie.	I like Bumblebee, he is cute.
区域(region)	2.95	Region of a movie.	Is it a Japanese movie?
上映时间(time)	2.02	Release time of a movie.	I remember it was in 1997, still at high school.
剧情(plot)	24.30	Plot of a movie.	What happened to him?
台词(line)	1.51	Line of a movie.	That line becomes a popular word that year.
奖项(awards)	1.17	Award of a movie.	It won a lot of awards that year.
票房(gross)	1.00	Gross of a movie.	I doubt if its gross will be satisfying.
评分(rating)	4.04	Rating of a movie.	It's clearly the best mivie I've watched.
资源(website)	1.25	Website of a movie.	You can watch it on iqiyi.com!
音乐(music)	2.19	Music in the movie.	Its theme song is "my heart will go on".
其他(other)	12.31	Other aspects.	I like her costume so much!

Table 11: Counts, descriptions and examples of the aspects. *Examples are shown only in translated English for space limit.*

B Human Evaluation

As for the four human evaluation metrics. The first two will focus only on the conversational backbones without considering domain knowledge. The second two will check if the responses can provide informative and correct responses powered by domain knowledge. The detailed definitions of them are:

A: 一直有听到《千与千寻》，却未曾真的看过这部电影	告知事实_类型; 告知感受_其他
B: 怎么不去看	其他
A: 话说你看了吗	其他
B: 看了，挺感动的	告知感受_其他
A: 主要讲什么的	请求事实_剧情
B: 一个小女孩，千寻，不畏艰难救了自己的父母和朋友	告知事实_角色; 告知事实_剧情
A: 海报里的小女孩看起来还很小	告知事实_角色
B: 所以啊，觉很她很勇敢	告知感受_其他
A: 听上去好像还挺好看的	告知感受_评分
B: 是啊，你快去看看吧	其他

A: 发现一部有点年代感的电影	告知感受_其他
B: 你说什么类型的	请求事实_类型
A: 特别反映现在的问题女人的地位	告知事实_剧情
B: 啥电影	请求事实_电影名
A: 菊豆，巩俐演的	告知事实_电影名 告知事实_演员名
B: 卧槽，她啊实力派啊自带气质真的	告知感受_演员名
A: 对啊我才看完	其他
B: 怎么样	请求感受_剧情
A: 有点点压抑，还是觉得我们现在的生活幸福	告知感受_剧情
B: 哈哈哈我也去看看	其他

A: 《蚁人》你看了没有	告知事实电影名
B: 蚁人绝对是最接地气的超级英雄！	告知事实_角色
A: 从小偷变英雄（因为是没有超能力吗？）	告知事实_其他
B: 好看死了！！每天炸纽约炸香港炸洛杉矶都看醉了！	告知感受_评分
A: 在玩具房里的打斗戏太有意思了！	告知事实_剧情; 告知感受_评分
B: 笑点也挺多的，看得我很开心。	告知感受_评分
A: Paul不是最帅的但绝对是最有味道的！	告知事实_角色
B: 就是最后十秒钟变星际穿越!真心吐槽	告知事实_剧情
A: 各种官方自带吐槽和自黑恶搞真是入戏不过三秒	告知感受_其他
B: 哈哈哈，确实是	其他

A: 你有没有看过多啦A梦的电影吗	告知事实_电影名
B: 你说的是哪一部？！	请求事实_电影名
A: 大雄的月球探险记	告知事实_电影名
B: 对啊，多啦A梦小时候最喜欢了，口袋里面什么都有	告知事实_剧情
A: 哈哈对，这里面他们去月球找玉兔，作者想象力很丰富！	告知事实_剧情
B: 后半部分其实有点融梗《星球大战》+《阿丽塔》	告知事实_电影名
A: 但也没有看出来有什么违和点啊	告知感受_评分
B: 恩呢，那倒是一直是萌物与科普，以及不变的友情。	告知感受_剧情
A: 最后胖虎把那个东西送给那个小孩的时候就哭了	告知事实_剧情
B: 你也一样嘛？	其他
A: 对啊对啊，就是蓝胖子的口袋东西太多了吧	告知事实剧情
B: 哈哈当然全都靠道具	告知事实剧情

1. Sensibleness: If the response makes sense given the dialogue context. *Specific domain knowledge is ignored*. For example, if the user asks "Who directed the movie Inception?", the bot replies "Steven Spielberg! I like him so much!!" The answer is wrong but still considered sensible given the context.

2. Engagement: If the response is engaging for the user to follow. The definition is generally similar to diversity, or interestingness metric used for open-domain chitchat. If the bot simply passively replies messages like "That's good" or "No idea", it will be marked as non-engaging.

3. Factuality: Factuality checks the information correctness in case the model fabricate wrong facts. This is orthogonal to the above metrics. The response can be factual but not sensible or the other way around. The same for Engagement.

4. Informativeness: Informativeness checks if the response contains new information about the movie. If the bot replies sth like "That's so interesting, I've never heard of that!". It is considered specific yet non-informative, because the bot provides no new information.

Intuitively a good closed-domain chatbot should balance well all these skills. Sensibleness plus engagement can evaluate its general performance as a chatbot. Factuality and informativeness can evaluate its capability at actively injecting appropriate domain knowledge into the dialogue.

All the three metrics are evaluated by three crow-workers each except for factuality. As evaluating factuality requires in-depth knowledge about one movie, it is quite difficult for random human evaluators to judge them. Even if we filter to only keep people who have watched a movie, it is hard to guarantee they can recall all the scenes in the movie. Therefore, the factuality check is only done by the person who performed this dialogue. In the static evaluation, it is evaluated by the annotator who produced the reference response. In the interactive evaluation, it is evaluated by the person who chat with the bot. It is nevertheless not accurate though. However, if the bot can cheat the human into believing its false information, it can also somehow be considered a "success". We provide examples for guiding the human evaluators in Table 12.

Table 13 shows some interactive examples with humans. We observe Mitsuku can XiaoIce perform decently in single-turn exchanges but strongly struggle at understanding multi-turn user intents. Most conversations stop at turn 4 and will not move on.

Conversation	Score
Input: Do you know who directed the movie Inception? Bot: I know! It was Christopher Nolan! I like him so much!	Sensibleness = 1; Engagement = 1 Factuality = 1, Informativeness = 1
Input: Do you know who directed the movie Inception? Bot: The movie is amazing to imagine about a dream within a dream.	Sensibleness = 0; Engagement = 1 Factuality = 1, Informativeness = 1
Input: Do you know who directed the movie Inception? Bot: Christopher Nolan.	Sensibleness = 1; Engagement = 0 Factuality = 1, Informativeness = 1
Input: Do you know who directed the movie Inception? Bot: I know! It was Steven Spielberg! I like him so much!	Sensibleness = 1; Engagement = 1 Factuality = 0, Informativeness = 1
Input: Do you know who directed the movie Inception? Bot: I don't remember. Do you like it?	Sensibleness = 1; Engagement = 1 Factuality = 1, Informativeness = 0
Input: Do you know who directed the movie Inception? Bot: Steven Spielberg.	Sensibleness = 1; Engagement = 0 Factuality = 0, Informativeness = 1

Table 12: Examples for human evaluation scores.

Mistsuku	XiaoIce	Ours
User: Do you know who directed the movie Inception?	*User*: 你喜欢看感人的电影吗	*User*: 你喜欢看感人的电影吗
Bot: No idea. I will guess at Steven Spielberg.	*Bot*: 我喜欢更加另类一些的	*Bot*: 喜欢啊，你有什么推荐吗
User: Maybe, it does look like his style.	*User*: 比如呢	*User*: 我不是药神很感人
Bot: Interesting comparison.	*Bot*: 比如？	*Bot*: 确实是一部很良心的电影
User:have you watched any movies recently?	*User*: 泰坦尼克号主角是谁	*User*: 确实，徐峥演的也好
Bot: My favourite movie is Terminator. Have you seen it?	*Bot*:这我还真知道，莱昂纳多·迪卡普里奥主演的	*Bot*: 徐峥的演绎一直很走心
User:not really, what is it about?	*User*:那是谁导演了呢	*User*:不过我最喜欢的还是里面的小黄毛
Bot:If not that, what?	*Bot*:你想让我说什么？	*Bot*: 小黄毛好难啊，太不容易了
User: I like the Avengers series so much	*User*:你看过大话西游吗	*User*:我看电影的时候他死那段给我哭完了
Bot: Yes. My favourite is Iron Man, as he's a kind of robot. You like Avengers.	*Bot*: 你可以再说一遍吗	*Bot*: 看完你有什么感想呢
User: Exactly! Can't accept he died in the end, sigh		*User*:最大的病就是穷病啊
Bot:You and I are on the same wavelength. I often wonder if it can.		*Bot*: 确实是这样，希望我们能平安
		User:平安是最大的幸福了
		Bot: 我觉得还行

Table 13: Sampled conversations from the interactive evaluation.