KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation

Hao Zhou*, **Chujie Zheng***, Kaili Huang, Minlie Huang, Xiaoyan Zhu

CoAI Group, DCST, Tsinghua University

About Me

- Chujie Zheng (郑楚杰)
- To be a Ph.D student this autumn in CoAI Group, THU
- Supervisor: Minlie Huang
- Got my B.Sc. in Dept. of Physics, THU

- Homepage: https://chujiezheng.github.io
- Contact: zcj16@tsinghua.org.cn

Outline

- Introduction
- Overview
- Dataset Collection
- Experiments
- Conclusion

Outline

- Introduction
  - Background
  - Motivation
- Overview
- Dataset Collection
- Experiments
- Conclusion

Background

- Background knowledge is crucial to dialog systems
  - For task-oriented, it (slot-value pair) provides essential info for QA & recommendation
  - For open-domain, it helps generate more informative and attractive responses
  - Structured knowledge graphs or unstructured texts

Background

- Existing open-domain dialogue corpora
  - Collect related external knowledge based on the context
    - Label the knowledge annotations using NER, string match, artificial scoring, and filtering rules
    - Mismatches introduce noises
  - Construct dialogues from scratch with human annotators
    - Maybe lack turn-level annotations
    - Constrained to 1-2 topics or lack of topic relations: limit modeling diversified topic transition and knowledge planning

Motivation

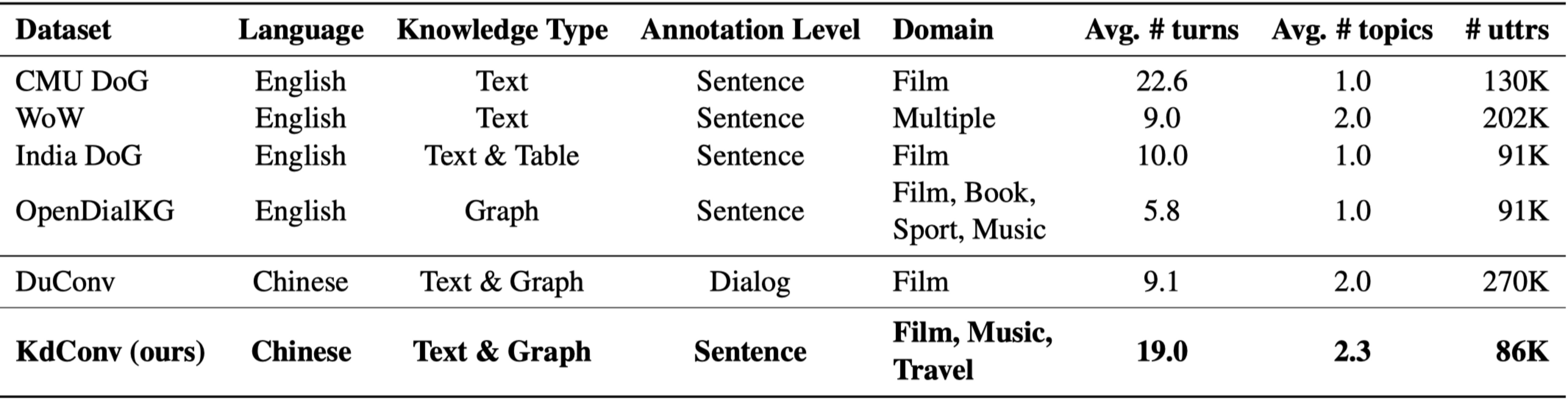
- Lack of dialog data on multiple topics with knowledge annotations
- Existing knowledge-grounded datasets have limitations in modeling knowledge interactions, such as topic transition and knowledge planning

Outline

- Introduction
- Overview
  - Comparison
  - Example Data
- Dataset Collection
- Experiments
- Conclusion

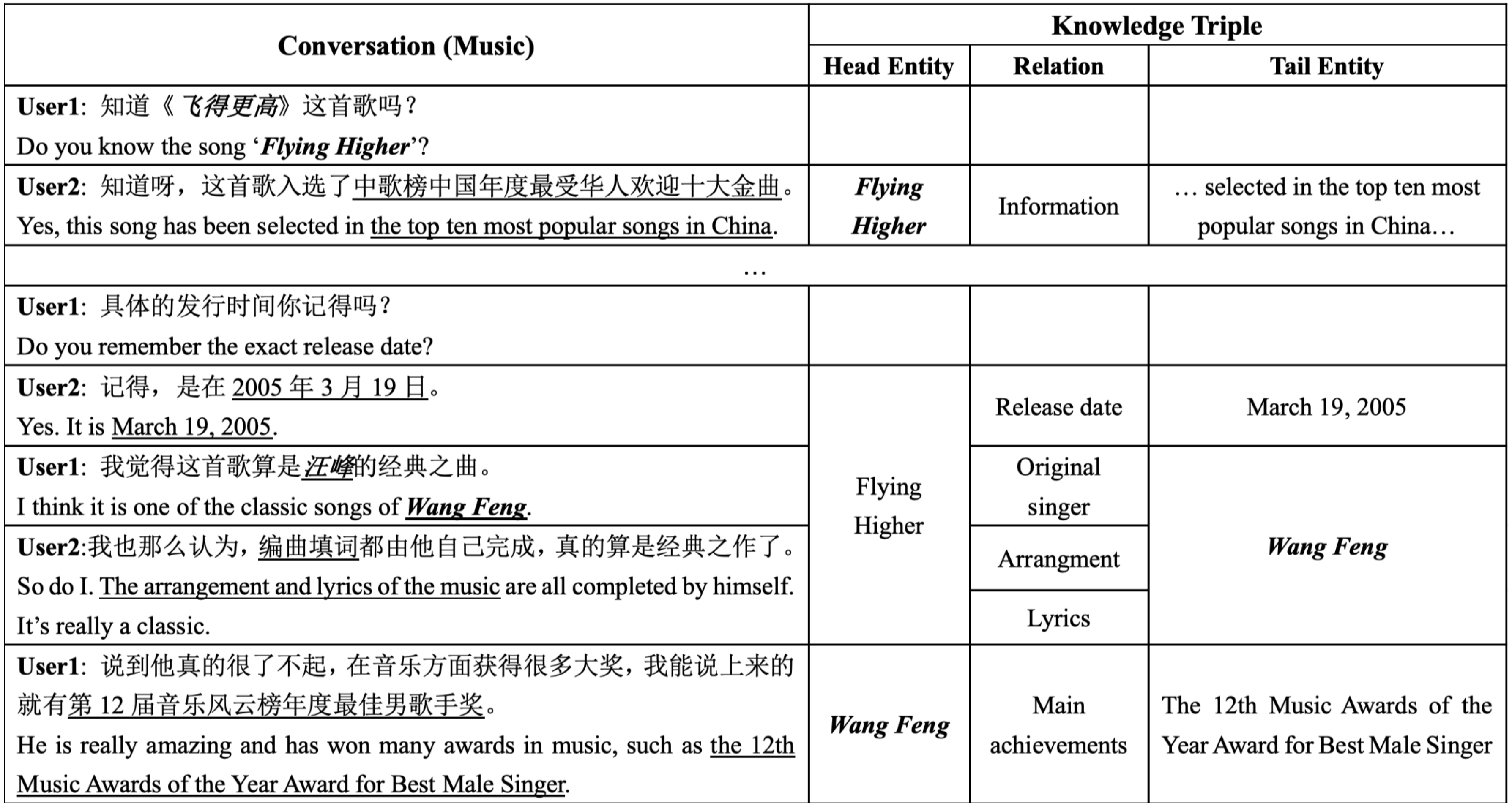
Overview

- Comparison



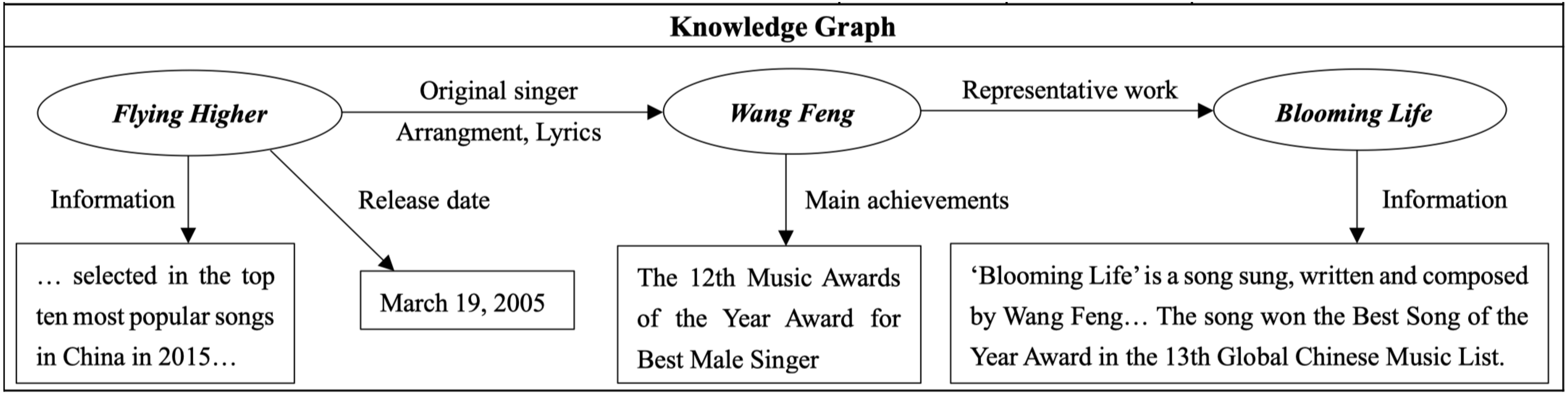
Dataset	Language	Knowledge Type	Annotation Level	Domain	Avg. # turns	Avg. # topics	# uttrs
CMU DoG	English	Text	Sentence	Film	22.6	1.0	130K
WoW	English	Text	Sentence	Multiple	9.0	2.0	202K
India DoG	English	Text & Table	Sentence	Film	10.0	1.0	91K
OpenDialKG	English	Graph	Sentence	Film, Book, Sport, Music	5.8	1.0	91K
DuConv	Chinese	Text & Graph	Dialog	Film	9.1	2.0	270K
KdConv (ours)	**Chinese**	**Text & Graph**	**Sentence**	**Film, Music, Travel**	**19.0**	**2.3**	**86K**

Overview



Conversation (Music)	Knowledge Triple		
	Head Entity	Relation	Tail Entity
User1：知道《*飞得更高*》这首歌吗？ Do you know the song ‘***Flying Higher***’?			
User2：知道呀，这首歌入选了中歌榜中国年度最受华人欢迎十大金曲。 Yes, this song has been selected in the top ten most popular songs in China.	***Flying Higher***	Information	… selected in the top ten most popular songs in China…
…			
User1：具体的发行时间你记得吗？ Do you remember the exact release date?			
User2：记得，是在 2005 年 3 月 19 日。 Yes. It is March 19, 2005.	Flying Higher	Release date	March 19, 2005
User1：我觉得这首歌算是*汪峰*的经典之曲。 I think it is one of the classic songs of ***Wang Feng***.		Original singer	***Wang Feng***
User2:我也那么认为，编曲填词都由他自己完成，真的算是经典之作了。 So do I. The arrangement and lyrics of the music are all completed by himself. It’s really a classic.		Arrangement Lyrics	
User1：说到他真的很了不起，在音乐方面获得很多大奖，我能说上来的就有第 12 届音乐风云榜年度最佳男歌手奖。 He is really amazing and has won many awards in music, such as the 12th Music Awards of the Year Award for Best Male Singer.	***Wang Feng***	Main achievements	The 12th Music Awards of the Year Award for Best Male Singer

Overview



Knowledge Graph

Flying Higher

Original singer

Arrangment, Lyrics

Wang Feng

Representative work

Blooming Life

Information

Release date

Main achievements

Information

… selected in the top ten most popular songs in China in 2015…

March 19, 2005

The 12th Music Awards of the Year Award for Best Male Singer

'Blooming Life' is a song sung, written and composed by Wang Feng… The song won the Best Song of the Year Award in the 13th Global Chinese Music List.

Outline

- Introduction
- Overview
- Dataset Collection
  - Knowledge Graph Construction
  - Dialogue Collection
  - Statistics
- Experiments
- Conclusion

Dataset Collection

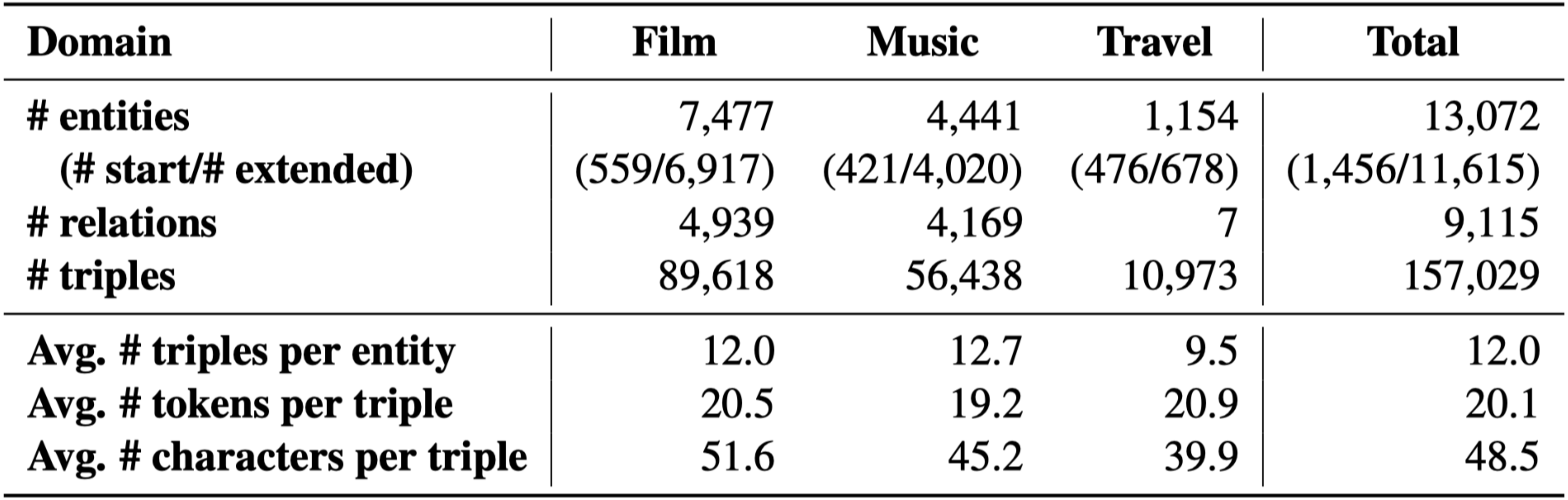
- Knowledge Graph Construction
  - Reduce the range of the domain-specific knowledge by crawling the most popular films and film stars, music and singers, and attractions as start entities, from several related websites (douban, qunar, etc.)
  - Filter the start entities which have few knowledge triples in XLORE (a large-scale English-Chinese bilingual knowledge graph)

Dataset Collection

- Knowledge Graph Construction
  - Retrieve their neighbor entities within three hops from XLORE
    - For the travel domain, the knowledge graph was crawled only from the Web, because XLORE provides little knowledge for start entities
  - Merge these entities and relations into a domain-specific knowledge graph

Dataset Collection

- Knowledge Graph Construction



Domain	Film	Music	Travel	Total
# entities	7,477	4,441	1,154	13,072
(# start/# extended)	(559/6,917)	(421/4,020)	(476/678)	(1,456/11,615)
# relations	4,939	4,169	7	9,115
# triples	89,618	56,438	10,973	157,029
Avg. # triples per entity	12.0	12.7	9.5	12.0
Avg. # tokens per triple	20.5	19.2	20.9	20.1
Avg. # characters per triple	51.6	45.2	39.9	48.5

Dataset Collection

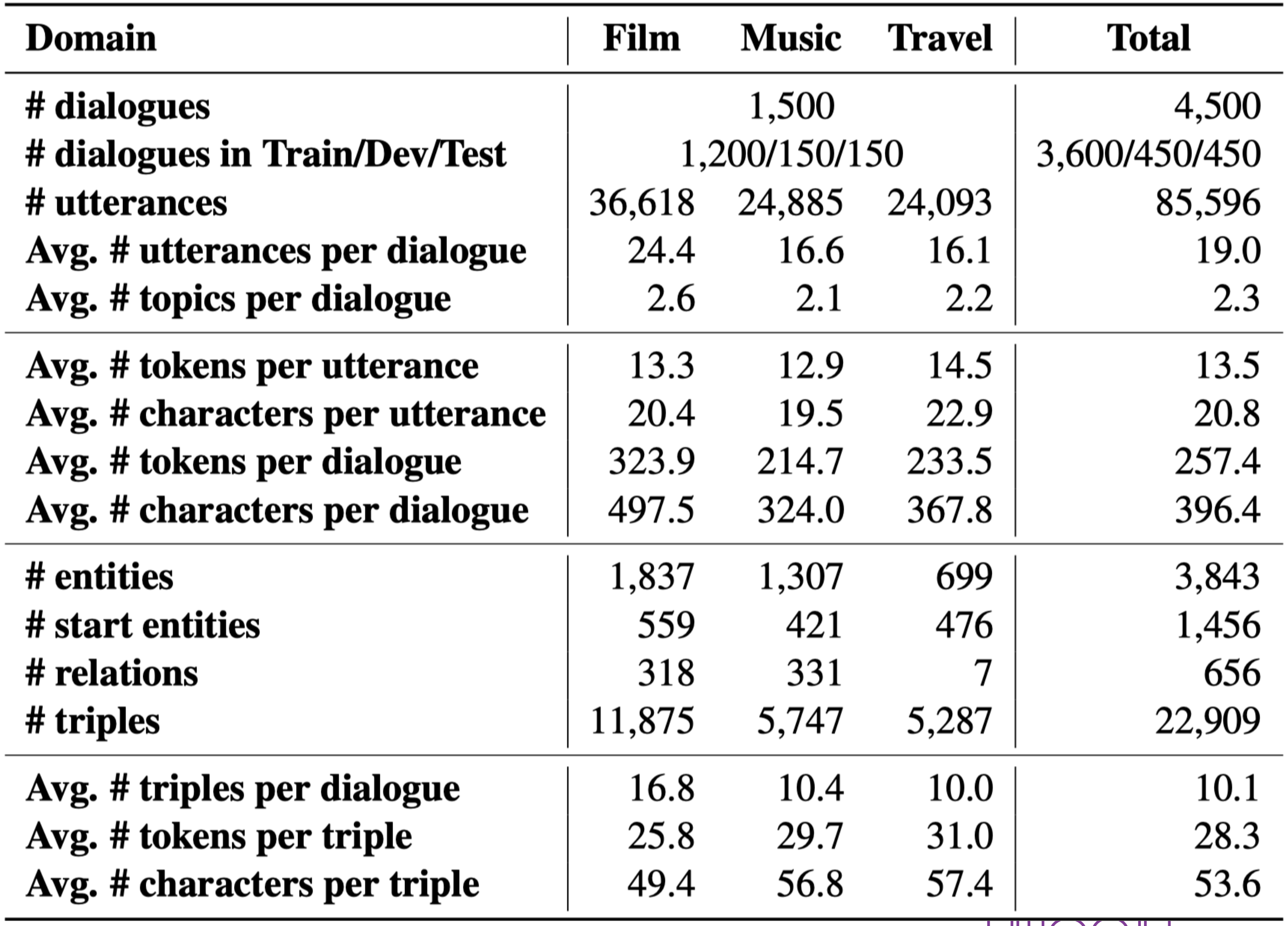
- Dialogue Collection
  - Recruit crowdsourced annotators to generate multi-turn without any pre-defined goals or constraints
  - During the conversation, two speakers both had access to the knowledge graph
  - Annotators were also required to record the related knowledge triples

Dataset Collection

- Dialogue Collection
  - Annotators were instructed to start the conversation based on start entities, and they were also encouraged to shift the topic of the conversation to other entities
  - Filter out low-quality dialogues, which contain grammatical errors, inconsistencies of knowledge facts, etc.

Dataset Collection

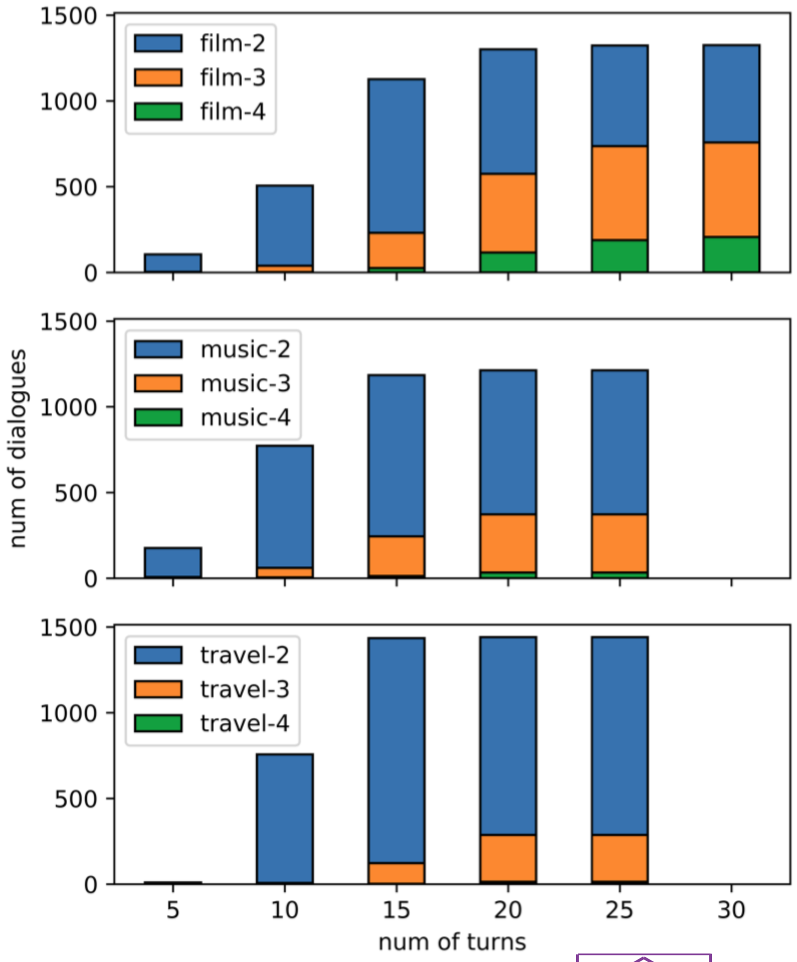
- Statistics
  - Film vs. music/travel
    - # turns
    - # topics
    - # triples



Domain	Film	Music	Travel	Total
# dialogues	1,500			4,500
# dialogues in Train/Dev/Test	1,200/150/150			3,600/450/450
# utterances	36,618	24,885	24,093	85,596
Avg. # utterances per dialogue	24.4	16.6	16.1	19.0
Avg. # topics per dialogue	2.6	2.1	2.2	2.3
Avg. # tokens per utterance	13.3	12.9	14.5	13.5
Avg. # characters per utterance	20.4	19.5	22.9	20.8
Avg. # tokens per dialogue	323.9	214.7	233.5	257.4
Avg. # characters per dialogue	497.5	324.0	367.8	396.4
# entities	1,837	1,307	699	3,843
# start entities	559	421	476	1,456
# relations	318	331	7	656
# triples	11,875	5,747	5,287	22,909
Avg. # triples per dialogue	16.8	10.4	10.0	10.1
Avg. # tokens per triple	25.8	29.7	31.0	28.3
Avg. # characters per triple	49.4	56.8	57.4	53.6

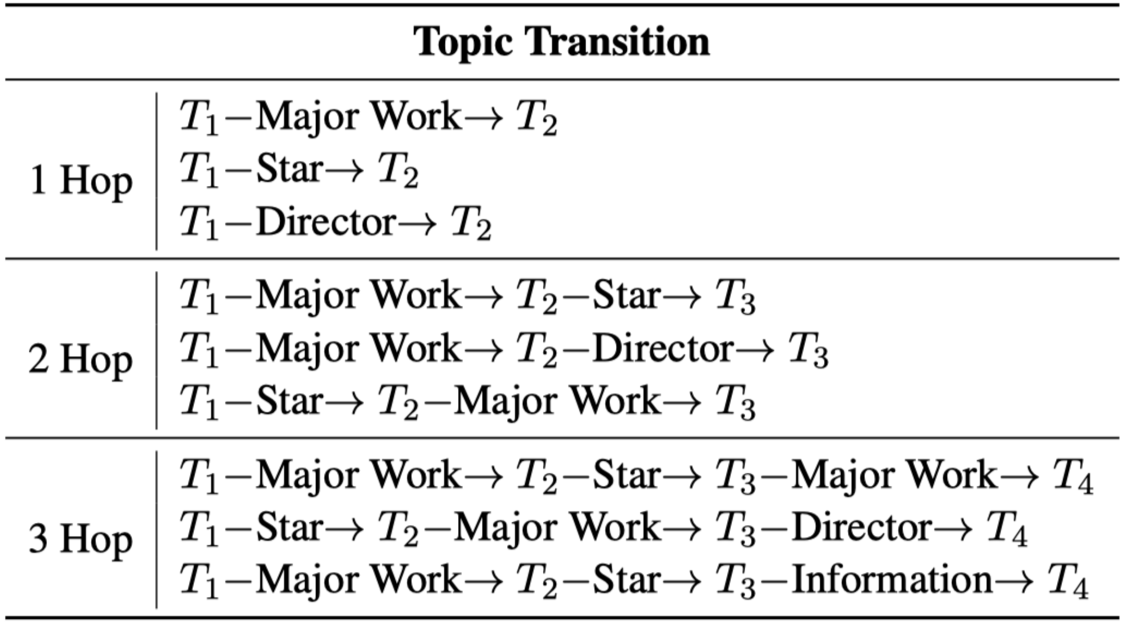
Dataset Collection

- Discussing multiple topics in depth usually requires a conversation having enough number of turns
  - A short conversation may not have natural transition between multiple topics



Dataset Collection

- Topic Transition
  - Diverse and complex
  - Suitable for the research of knowledge planning
  - “−Info→” : people prefer to shift the topic according to the structured relations rather than unstructured texts



Topic Transition	
1 Hop	T_1−Major Work→ T_2
	T_1−Star→ T_2
	T_1−Director→ T_2
2 Hop	T_1−Major Work→ T_2−Star→ T_3
	T_1−Major Work→ T_2−Director→ T_3
	T_1−Star→ T_2−Major Work→ T_3
3 Hop	T_1−Major Work→ T_2−Star→ T_3−Major Work→ T_4
	T_1−Star→ T_2−Major Work→ T_3−Director→ T_4
	T_1−Major Work→ T_2−Star→ T_3−Information→ T_4

Outline

- Introduction
- Overview
- Dataset Collection
- Experiments
  - Benchmark Models
  - Automatic/Manual Evaluation
  - Case Study
- Conclusion

Experiments

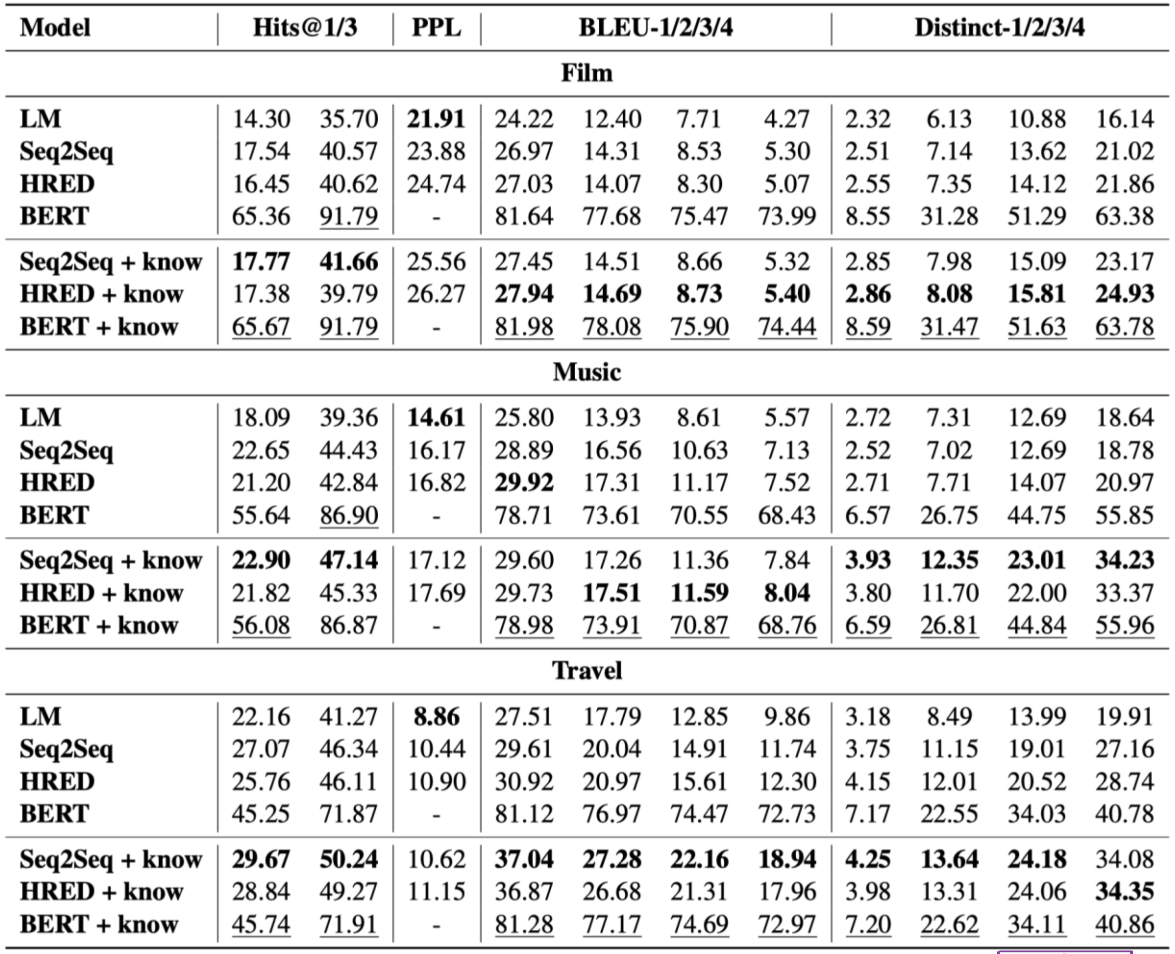
- Benchmark Models
  - Generation-based Models
    - Language Model (Bengio et al., 2003)
    - Seq2Seq (Sutskever et al., 2014)
    - HRED (Serban et al., 2016)
  - Retrieval-based Model: BERT (NSP) (Devlin et al., 2019)
  - Knowledge-aware Models
    - Key-Value Memory Module (Miller et al., 2016)

Experiments

- Advertisement
  - CoTK (Conversational Toolkit): An open-source toolkit for fast development and fair evaluation of language generation
    - Predefined evaluation suites, test models with popular and standard metrics
  - Paper: https://arxiv.org/abs/2002.00583
  - GitHub: https://github.com/thu-coai/cotk

Experiments

- Automatic Evaluation
  - Metrics
    - Hits@n
    - PPL
    - BLEU
    - Distinct
  - Results



Model	Hits@1/3		PPL	BLEU-1/2/3/4				Distinct-1/2/3/4			
					Film						
LM	14.30	35.70	**21.91**	24.22	12.40	7.71	4.27	2.32	6.13	10.88	16.14
Seq2Seq	17.54	40.57	23.88	26.97	14.31	8.53	5.30	2.51	7.14	13.62	21.02
HRED	16.45	40.62	24.74	27.03	14.07	8.30	5.07	2.55	7.35	14.12	21.86
BERT	65.36	91.79	-	81.64	77.68	75.47	73.99	8.55	31.28	51.29	63.38
Seq2Seq + know	**17.77**	**41.66**	25.56	27.45	14.51	8.66	5.32	2.85	7.98	15.09	23.17
HRED + know	17.38	39.79	26.27	**27.94**	**14.69**	**8.73**	**5.40**	**2.86**	**8.08**	**15.81**	**24.93**
BERT + know	65.67	91.79	-	81.98	78.08	75.90	74.44	8.59	31.47	51.63	63.78
					Music						
LM	18.09	39.36	**14.61**	25.80	13.93	8.61	5.57	2.72	7.31	12.69	18.64
Seq2Seq	22.65	44.43	16.17	28.89	16.56	10.63	7.13	2.52	7.02	12.69	18.78
HRED	21.20	42.84	16.82	**29.92**	17.31	11.17	7.52	2.71	7.71	14.07	20.97
BERT	55.64	86.90	-	78.71	73.61	70.55	68.43	6.57	26.75	44.75	55.85
Seq2Seq + know	**22.90**	**47.14**	17.12	29.60	17.26	11.36	7.84	**3.93**	**12.35**	**23.01**	**34.23**
HRED + know	21.82	45.33	17.69	29.73	**17.51**	**11.59**	**8.04**	3.80	11.70	22.00	33.37
BERT + know	56.08	86.87	-	78.98	73.91	70.87	68.76	6.59	26.81	44.84	55.96
					Travel						
LM	22.16	41.27	**8.86**	27.51	17.79	12.85	9.86	3.18	8.49	13.99	19.91
Seq2Seq	27.07	46.34	10.44	29.61	20.04	14.91	11.74	3.75	11.15	19.01	27.16
HRED	25.76	46.11	10.90	30.92	20.97	15.61	12.30	4.15	12.01	20.52	28.74
BERT	45.25	71.87	-	81.12	76.97	74.47	72.73	7.17	22.55	34.03	40.78
Seq2Seq + know	**29.67**	**50.24**	10.62	**37.04**	**27.28**	**22.16**	**18.94**	**4.25**	**13.64**	**24.18**	34.08
HRED + know	28.84	49.27	11.15	36.87	26.68	21.31	17.96	3.98	13.31	24.06	**34.35**
BERT + know	45.74	71.91	-	81.28	77.17	74.69	72.97	7.20	22.62	34.11	40.86

Experiments

Manual Evaluation

Metrics (0-2)

- Fluency: whether the response is fluent and natural
- Coherence: whether a response is relevant and coherent to the context and knowledge

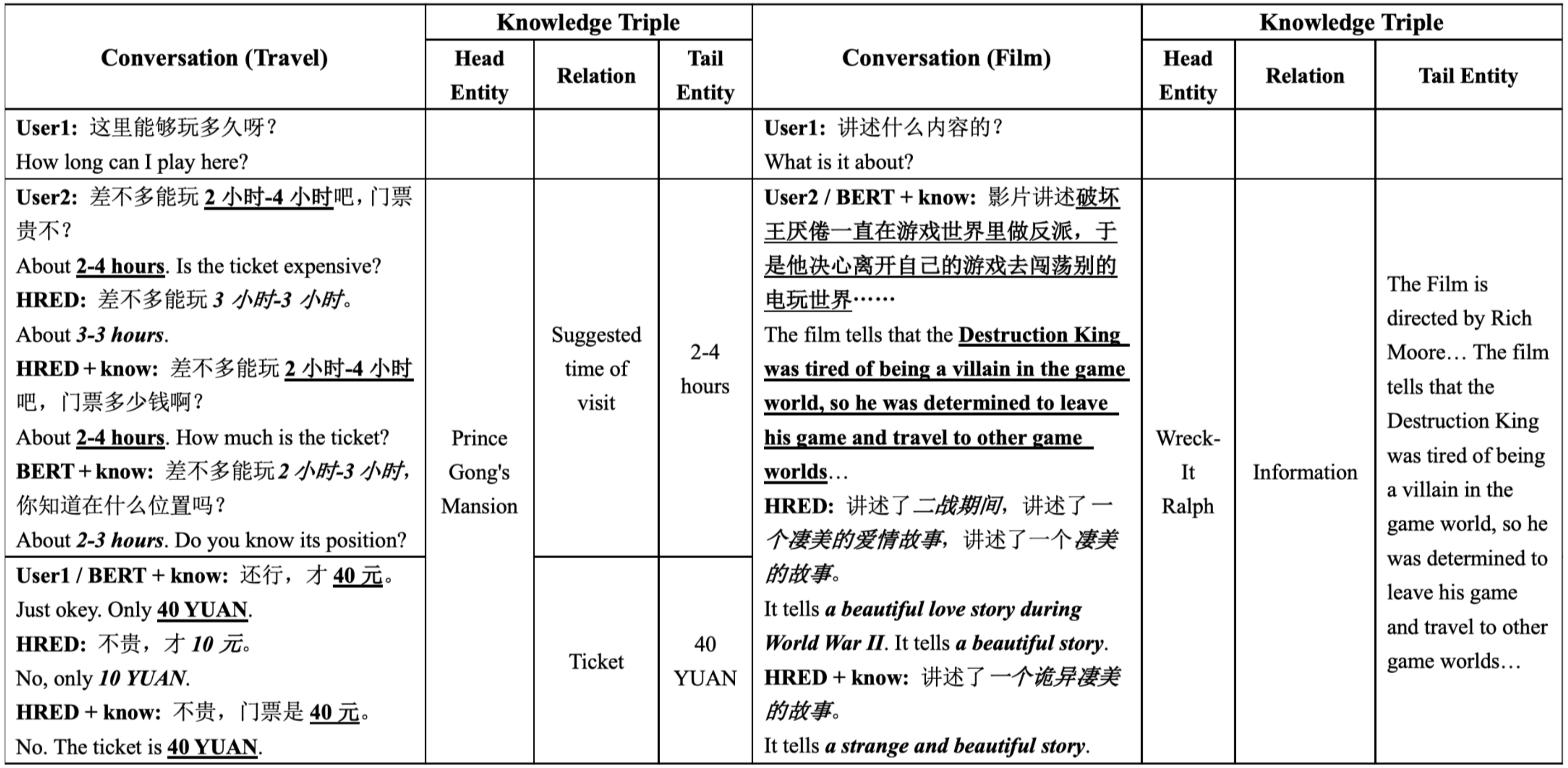
Annotation Statistics

- 500 samples each domain, 3 annotators
- Fleiss’ kappa: from 0.37 to 0.74
- 3/3 agreement: from 68.14% to 81.33%

Model	**Fluency**	**Coherence**
Film \ κ	0.50	0.61
HRED	1.64	1.19
HRED + know	*1.78*	*1.28*
BERT + know	**2.00**	**1.79**
Music \ κ	0.37	0.57
HRED	1.90	1.30
HRED + know	1.86	1.36
BERT + know	**2.00**	**1.80**
Travel \ κ	0.55	0.74
HRED	1.77	1.10
HRED + know	1.78	*1.31*
BERT + know	**2.00**	**1.76**

Experiments

Case Study



Conversation (Travel)	Knowledge Triple			Conversation (Film)	Knowledge Triple		
	Head Entity	Relation	Tail Entity		Head Entity	Relation	Tail Entity
User1: 这里能够玩多久呀？ How long can I play here?				**User1:** 讲述什么内容的？ What is it about?			
User2: 差不多能玩 **2 小时-4 小时**吧，门票贵不？ About **2-4 hours**. Is the ticket expensive? **HRED:** 差不多能玩 ***3 小时-3 小时***。 About ***3-3 hours***. **HRED + know:** 差不多能玩 **2 小时-4 小时**吧，门票多少钱啊？ About **2-4 hours**. How much is the ticket? **BERT + know:** 差不多能玩 ***2 小时-3 小时***，你知道在什么位置吗？ About ***2-3 hours***. Do you know its position?	Prince Gong's Mansion	Suggested time of visit	2-4 hours	**User2 / BERT + know:** 影片讲述破坏王厌倦一直在游戏世界里做反派，于是他决心离开自己的游戏去闯荡别的电玩世界…… The film tells that the **Destruction King was tired of being a villain in the game world, so he was determined to leave his game and travel to other game worlds**… **HRED:** 讲述了*二战期间*，讲述了一*个凄美的爱情故事*，讲述了一个*凄美的故事*。 It tells ***a beautiful love story during World War II***. It tells ***a beautiful story***. **HRED + know:** 讲述了 *一个诡异凄美的故事*。 It tells ***a strange and beautiful story***.	Wreck-It Ralph	Information	The Film is directed by Rich Moore… The film tells that the Destruction King was tired of being a villain in the game world, so he was determined to leave his game and travel to other game worlds…
User1 / BERT + know: 还行，才 **40 元**。 Just okey. Only **40 YUAN**. **HRED:** 不贵，才 ***10 元***。 No, only ***10 YUAN***. **HRED + know:** 不贵，门票是 **40 元**。 No. The ticket is **40 YUAN**.		Ticket	40 YUAN				

Outline

- Introduction
- Overview
- Dataset Collection
- Experiments
- Conclusion

Conclusion

- Propose a Chinese multi-domain corpus for **K**nowledge-**d**riven **Con**versation, KdConv, which contains 86K utterances and 4.5K dialogues, with an average number of 19.0 turns
- Extensive experiments demonstrate that the models can be enhanced by introducing knowledge, whereas there is still much room in knowledge-grounded conversation modeling for future work

Conclusion

- Competition
    - SMP2020-ECDT (The Evaluation of Chinese Human-Computer Dialogue Technology, Task 2)
    - The participators need to retrieve relevant triples from the domain-specific graph

Thanks for your attention

Paper: https://arxiv.org/abs/2004.04100

GitHub: https://github.com/thu-coai/KdConv