

# Related works on Conversational Reasoning

Qintong Li

July 30<sup>th</sup> 2020

# Related Paper list

- OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs (Facebook. ACL 2019)
- KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven (THU. ACL 2020)
- Commonsense Transformers for Automatic Knowledge Graph Construction (UW. ACL 2019)
- MuTual: A Dataset for Multi-Turn Dialogue Reasoning (ZJU etc. ACL 2020)
- ASER: A Large-scale Eventuality Knowledge Graph (HKUST. WWW 2020)
- Guided Generation of Cause and Effect (HIT JHU. IJCAI 2020)

# OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs

Seungwhan Moon, Pararth Shah, Anuj Kumar, Rajen Subba

Facebook Conversational AI

{shanemoon, pararths, anujk, rasubba@}fb.com

ACL 2019

# Motivation

Key elements:

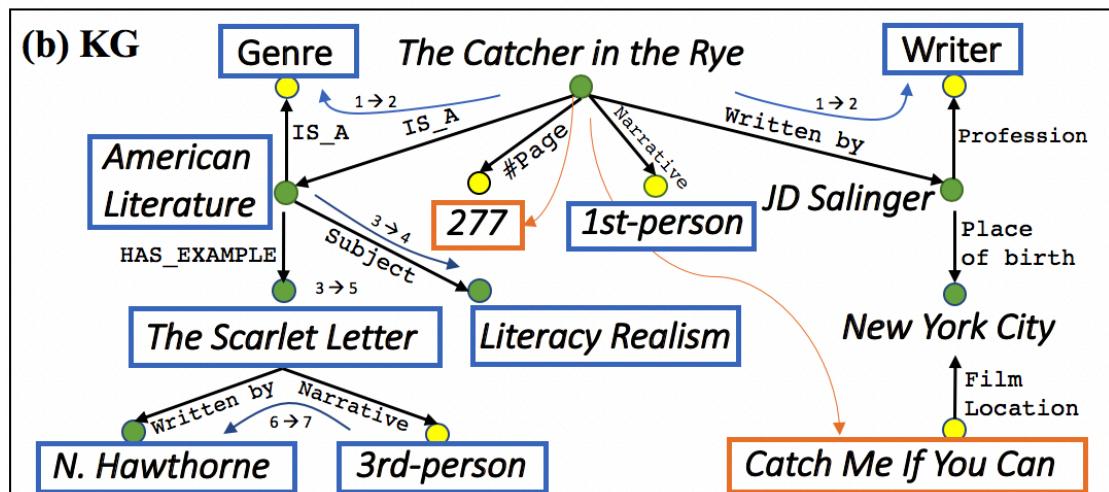
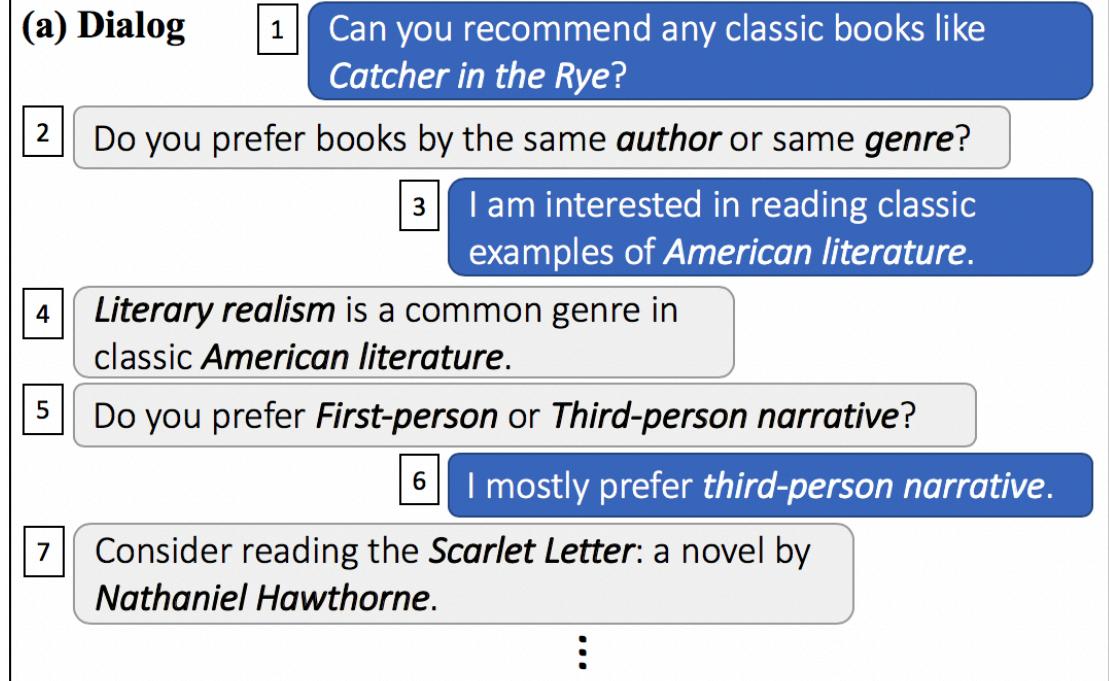
- Understand conversational contexts;
- Respond naturally by introducing relevant entities and attributes.

Core challenges:

- Domain-agnostic;
- Scalable prediction that follows natural conceptual threads.

A data-driven conversational reasoning model.

- Walkable degree of each entity varies by dialog contexts and domains;
- Pruning the search space for entities is a crucial step in operating knowledge-augmented dialog systems at scale.

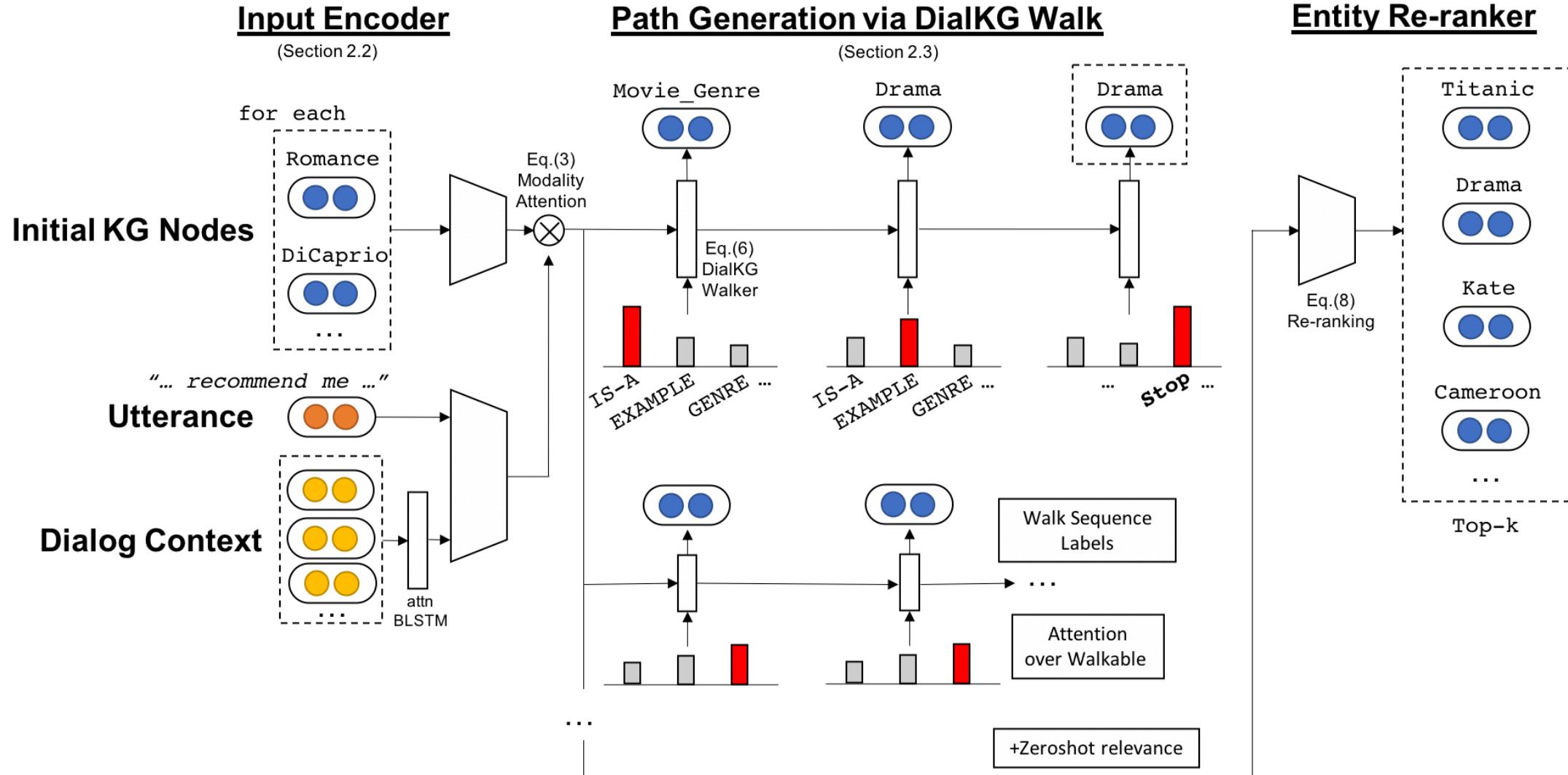


Non-ideal entities.

# Overview

- They propose a new model that can learn natural *knowledge paths* among entities mentioned over dialog contexts, and *reason* grounded on a large commonsense KG (Freebase (Bast et al., 2014)).
- They collect a new human-to-human multi-turn dialogs dataset (91K utterances across 15K dialog sessions) where each utterance is annotated with mentioned entities and factual connections.
- They completely ground dialogs in a large-scale common-fact KG, allowing for **domain-agnostic conversational reasoning**. Extensive cross-domain and transfer learning evaluations demonstrate model's flexibility.

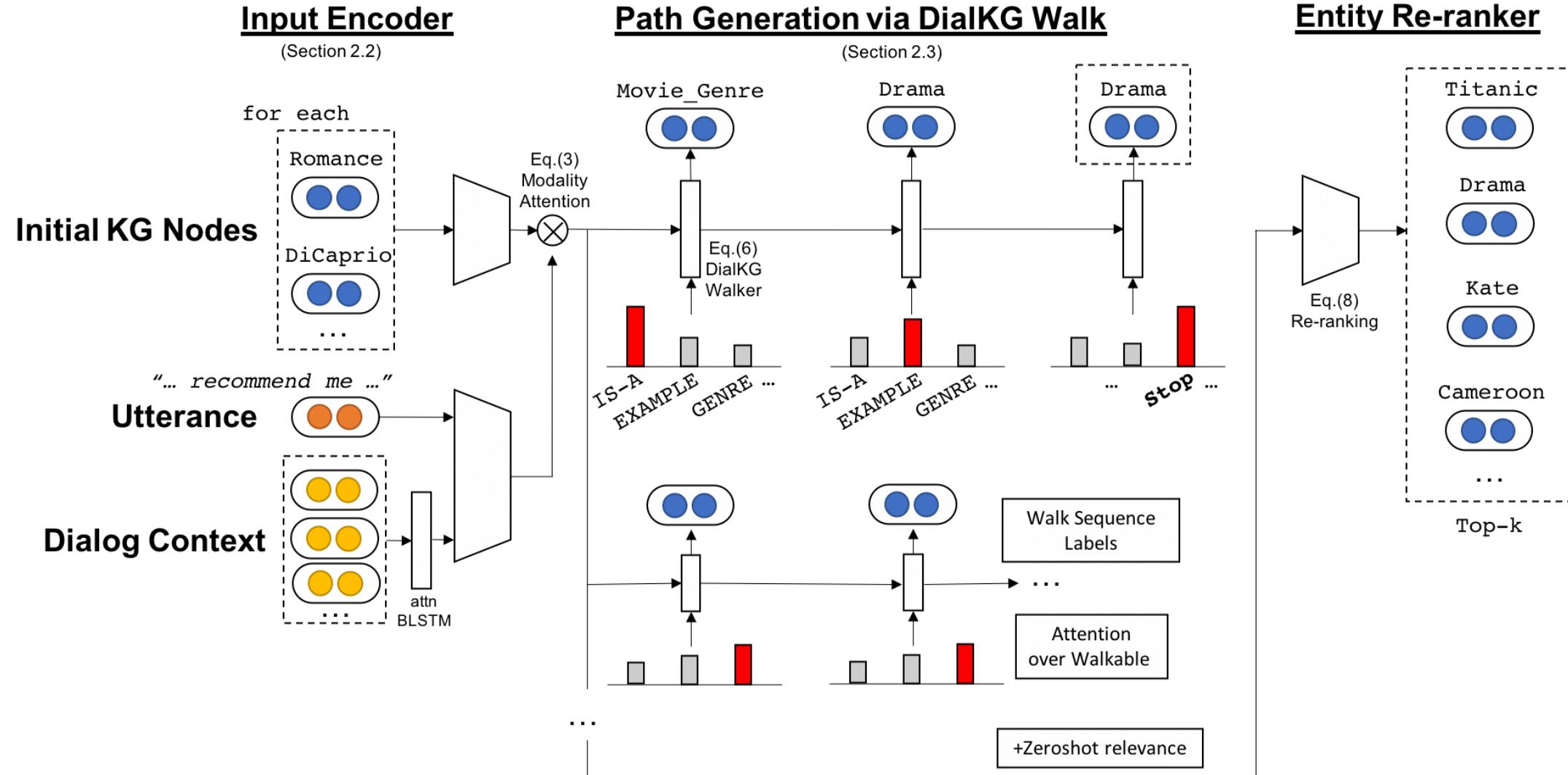
# Method



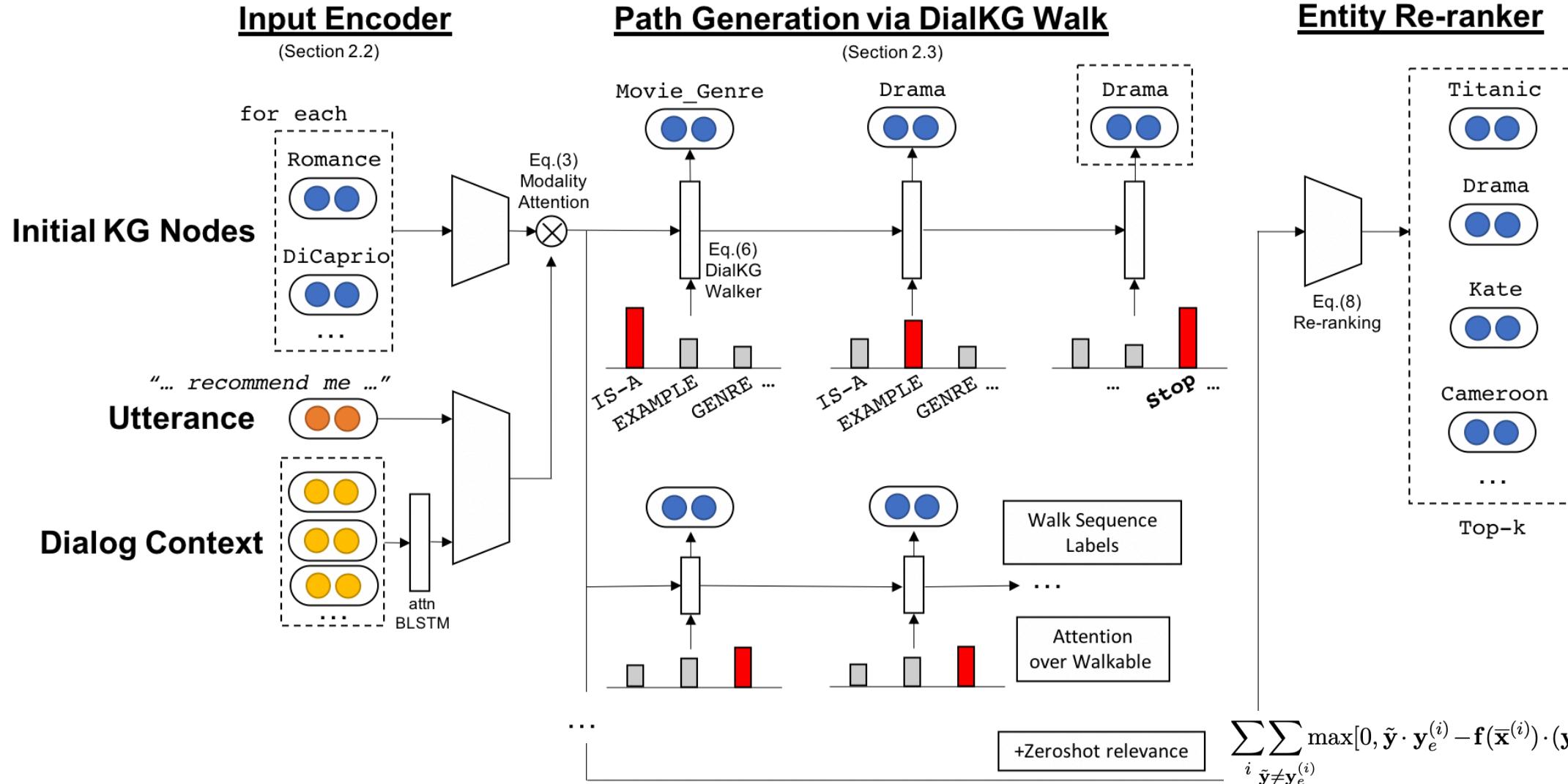
# Method

$$\min_{\mathbf{W}} \mathcal{L}_f(\mathbf{x}, \mathbf{y}_e; \mathbf{W}_f, \mathbf{W}_p) + \mathcal{L}_{\text{walk}}(\mathbf{x}, \mathbf{y}_p; \mathbf{W}_p)$$

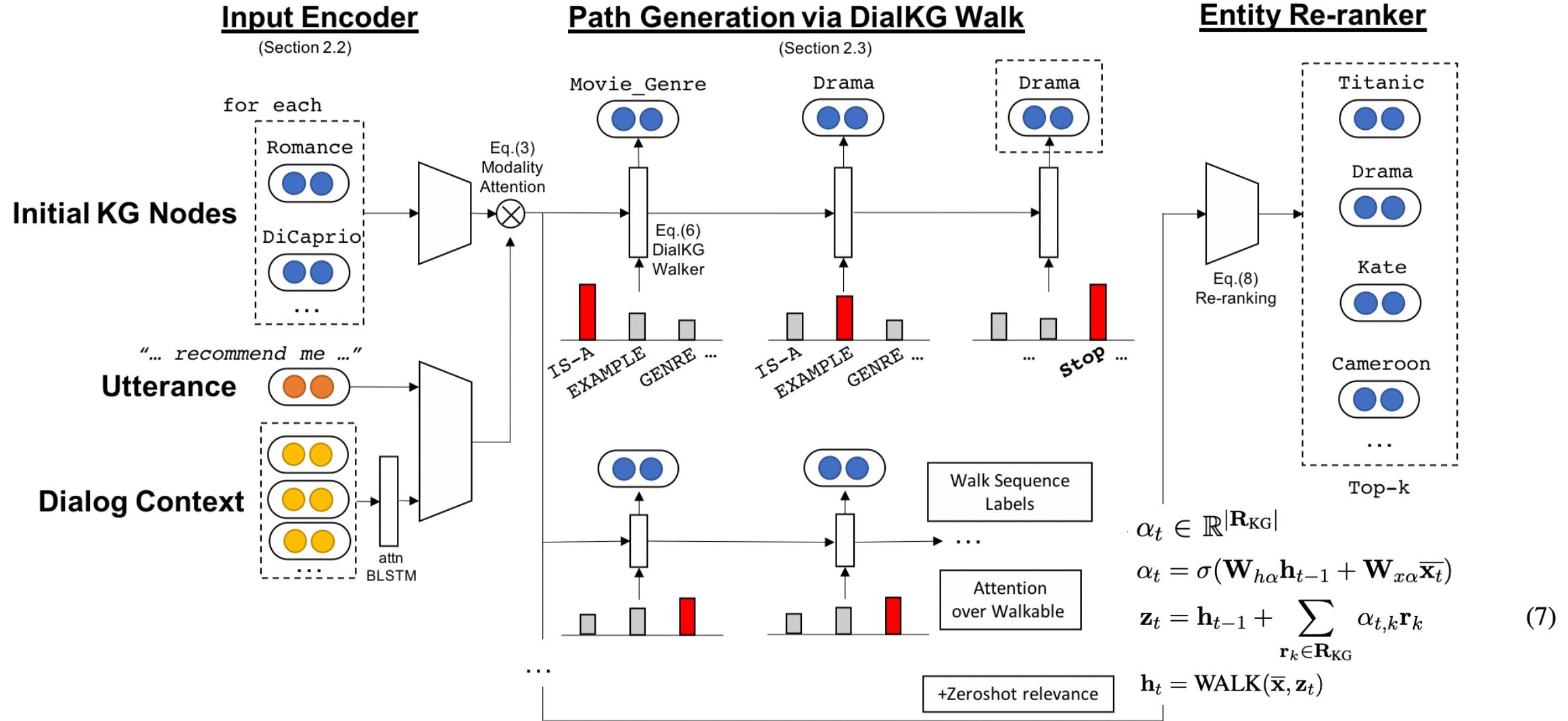
$$\mathcal{R}(\mathbf{W}): \text{regularization} \quad (4)$$



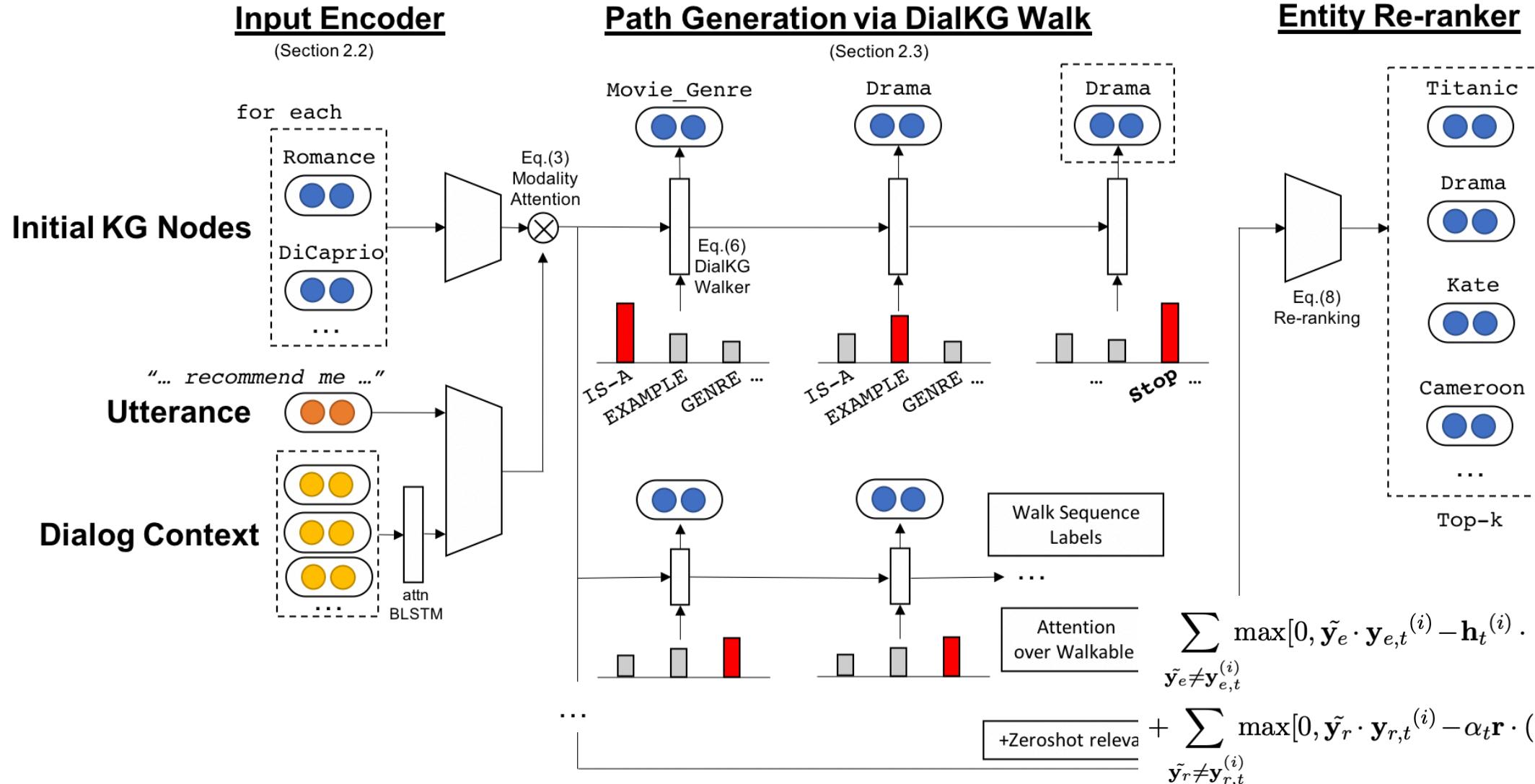
# Method



# Method



# Method



# Method



Adversarial Transfer Learning

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_f + \mathcal{L}_{\text{walk}} + \text{Entropy}(\sigma(\mathbf{W}_d \mathbf{x}), \mathbf{y}_d) \\ \mathbf{h}_t &= \text{WALK}([\bar{\mathbf{x}}; (\mathbf{W}_d \mathbf{x})], \mathbf{z}_t)\end{aligned}\quad (9)$$

# Experiments

Input	Model	Movie → Book					Movie → Music				
		r@1	3	5	10	25	r@1	3	5	10	25
E + S + D	seq2seq (Sutskever et al., 2014)	2.9	21.3	35.1	50.6	64.2	1.5	12.1	19.7	34.9	49.4
E + S	Tri-LSTM (Young et al., 2018)	2.3	17.9	29.7	44.9	61.0	1.9	8.7	12.9	25.8	44.4
E + S	Ext-ED (Parthasarathi and Pineau, 2018)	2.0	7.9	11.2	16.4	22.4	1.3	2.6	3.8	4.1	8.3
E	DialKG Walker (ablation)	8.2	15.7	22.8	31.8	48.9	4.5	<b>16.7</b>	21.6	25.8	33.0
E + S	DialKG Walker (ablation)	12.6	28.6	38.6	<b>54.1</b>	<b>65.6</b>	<b>6.0</b>	15.9	<b>22.8</b>	<b>33.0</b>	<b>47.5</b>
E + S + D	DialKG Walker ( <b>proposed</b> )	<b>13.5</b>	<b>28.8</b>	<b>39.5</b>	52.6	64.8	5.3	13.3	19.7	28.8	38.0

Table 3: Cross-domain (train/test on the different domain) response generation performance on the *OpenDialKG* dataset (metric: recall@ $k$ ). E: entities, S: sentence, D: dialog contexts.  
 (before masking). E: entities, S: sentence, D: dialog contexts.

# Experiments

1. Achieve the best performance especially for domains that are semantically close (*e.g.* movie and book);
2. transfers knowledge from a pre-trained source model via fine-tuning (hence requiring significantly less training resources), and effectively avoids “cold start”;
3. the DialKG model can quickly adapt to other new low-resource domains and improve upon the zero-shot cross-domain performance

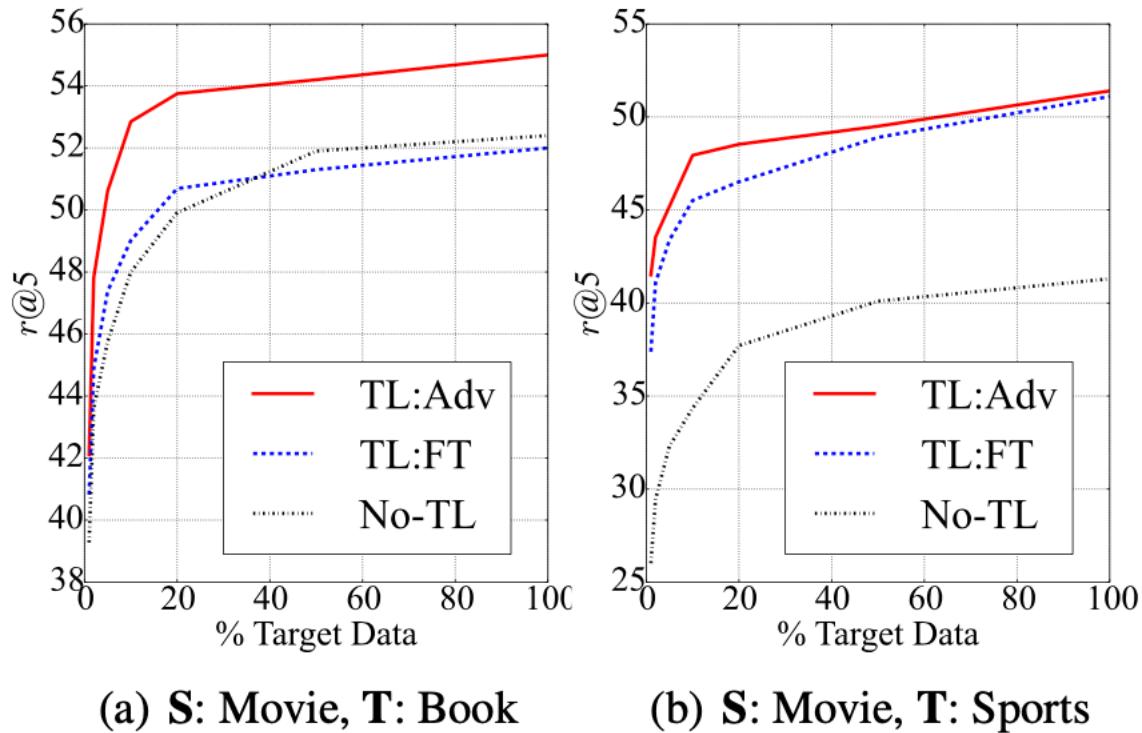


Figure 3: **Transfer learning** results ( $r@5$ ) of DialKG Walker at varying availability of target data with (a) Book and (b) Sports domains as a Target (Source: Movie). (TL:Adv): data transfer with adversarial discriminator for source and target domains, (TL:FT): model transfer with fine-tuning, (No-TL): target only.

# Summary

## ➤ Strength

- A new conversational reasoning model can navigate a large- scale, open-ended KG given conversational contexts.
- The new dataset provides a new way to study how conversational topics could jump across many different entities and KG paths within multi-turn dialog setting.
- Zeroshot relevance and transfer learning may help the domain-agnostic conversational reasoning.

## ➤ Weakness

- This parallel corpus of textual dialogs and corresponding KG walks is impractical. What if KG/dialogue dataset updates? Human annotation is time-consuming.
- They only consider the entities in the current turn.
- End-to-end generation of sentences (e.g. based on the retrieved entities) is not part of this study.

<https://github.com/facebookresearch/opendialkg> (data)

# KdConv Dataset

Hao Zhou, Chujia  
AI Group  
Beijing National  
tuxchow@gmail.com

Dataset
CMU DoG
WoW
India DoG
OpenDialKG
DuConv
<b>KdConv (ours)</b>

Table 1: Co

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 ' <i>Flying Higher</i> '?			
User2: 知道呀, 这首歌入选了 <u>中歌榜中国年度最受华人欢迎十大金曲</u> 。 Yes, this song has been selected in <u>the top ten most popular songs in China</u> .	<i>Flying Higher</i>	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 <u>2005年3月19日</u> 。 Yes. It is <u>March 19, 2005</u> .		Release date	March 19, 2005
User1: 我觉得这首歌算是 <u>汪峰</u> 的经典之曲。 I think it is one of the classic songs of <u>Wang Feng</u> .		Original singer	
User2: 我也那么认为, <u>编曲填词</u> 都由他自己完成, 真的算是经典之作。 So do I. <u>The arrangement and lyrics of the music</u> are all completed by himself. It's really a classic.		Arrangement	<i>Wang Feng</i>
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 <u>第12届音乐风云榜年度最佳男歌手奖</u> 。 He is really amazing and has won many awards in music, such as <u>the 12th Music Awards of the Year Award for Best Male Singer</u> .	<i>Wang Feng</i>	Main achievements	The 12th Music Awards of the Year Award for Best Male Singer
User1: 那他的歌曲除飞得更高, 你还喜欢哪首? So which song do you like besides 'Flying Higher'?			
User2: 再喜欢的就是《 <u>怒放的生命</u> 》这首歌了, 听的感觉特别好, 减压。 I like ' <u>Blooming Life</u> '. I feel great and decompression.	<i>Wang Feng</i>	Representative works	Fireworks, Brave Heart, Flying Higher, <i>Blooming Life</i> ...
User1: 啊, 这首歌我也很喜欢, 也都是由他自己 <u>作词作曲并演唱</u> 。 Oh, I like this song, too. He <u>wrote and sang</u> it by himself.			'Blooming Life' is a song sung, written and composed by Wang Feng...
User2: 是的, 该曲也获得了 <u>13届全球华语音乐榜中榜年度最佳歌曲奖</u> 。 Yes, and the song also won <u>the Best Song of the Year Award in the 13th Global Chinese Music List</u> .	<i>Blooming Life</i>	Information	The song won the Best Song of the Year Award in the 13th Global Chinese Music List.
Knowledge Graph			

Figure 1: An example in KdConv from the **music** domain. The underlined text is the related knowledge that is utilized in conversation. The **italic** text and circles are **topics** (refer to the distinct head entities in the knowledge triples and the central nodes with degree greater than 1 in the knowledge graph) in this dialogue.

ialogue  
e-driven  
u Conversational  
Jniversity  
chnology, China  
singhua.edu.cn

topics	# uttrs
0	130K
0	202K
0	91K
0	91K
0	270K
<b>3</b>	<b>86K</b>

gue corpora.

# Summary

- A Chinese version of OpenDialKG;
- This parallel corpus of textual dialogs and corresponding KG is impractical. What if related entities not exist? KG/dialogue dataset updates? Human annotation is time-consuming.

# COMET<sup>◊</sup>: Commonsense Transformers for Automatic Knowledge Graph Construction

**Antoine Bosselut ♦♦ ♠ Hannah Rashkin ♦♦ ♠ Maarten Sap ♦♦ ♠ Chaitanya Malaviya ♦  
Asli Celikyilmaz ♣ Yejin Choi ♦♦ ♠**

◊Allen Institute for Artificial Intelligence, Seattle, WA, USA

♠Paul G. Allen School of Computer Science & Engineering, Seattle, WA, USA  
♣Microsoft Research, Redmond, WA, USA

## Wikipedia (ground truth)

dewey & leboeuf llp was a global law firm , headquartered in new york city , that is now in bankruptcy . the firm 's leaders have been indicted for fraud for their role in allegedly cooking the company 's books to obtain loans while hiding the firm 's financial plight . the firm was formed in 2007 through the merger of dewey ballantine and leboeuf lamb orman & mcalpin .  
dewey & leboeuf was known for its corporate , insurance , litigation , tax and real estate law . after filing for bankruptcy , it employed over 1,000 lawyers in 26 offices around the world . its indebtedness became public . in the same period , many partners departed , an investigation began to investigate alleged false statements by firm chairman steven davis . as leboeuf 's offices began to enter administration in may 2012 . the firm filed for bankruptcy protection on march 6 , 2014 , the former chairman , chief financial officer and the executive director of finance were charged with grand larceny by the manhattan district attorney .

**Situation:** Speaker felt this when...

"I've been hearing noises around the house at night"

**Conversation:**

**Speaker:** I've been hearing some strange noises around the house at night.

**Listener:** oh no! That's scary! What do you think it is?

**Speaker:** I don't know, that's what's making me anxious.

**Listener:** I'm sorry to hear that. I wish I could help you figure it out

# Challenge

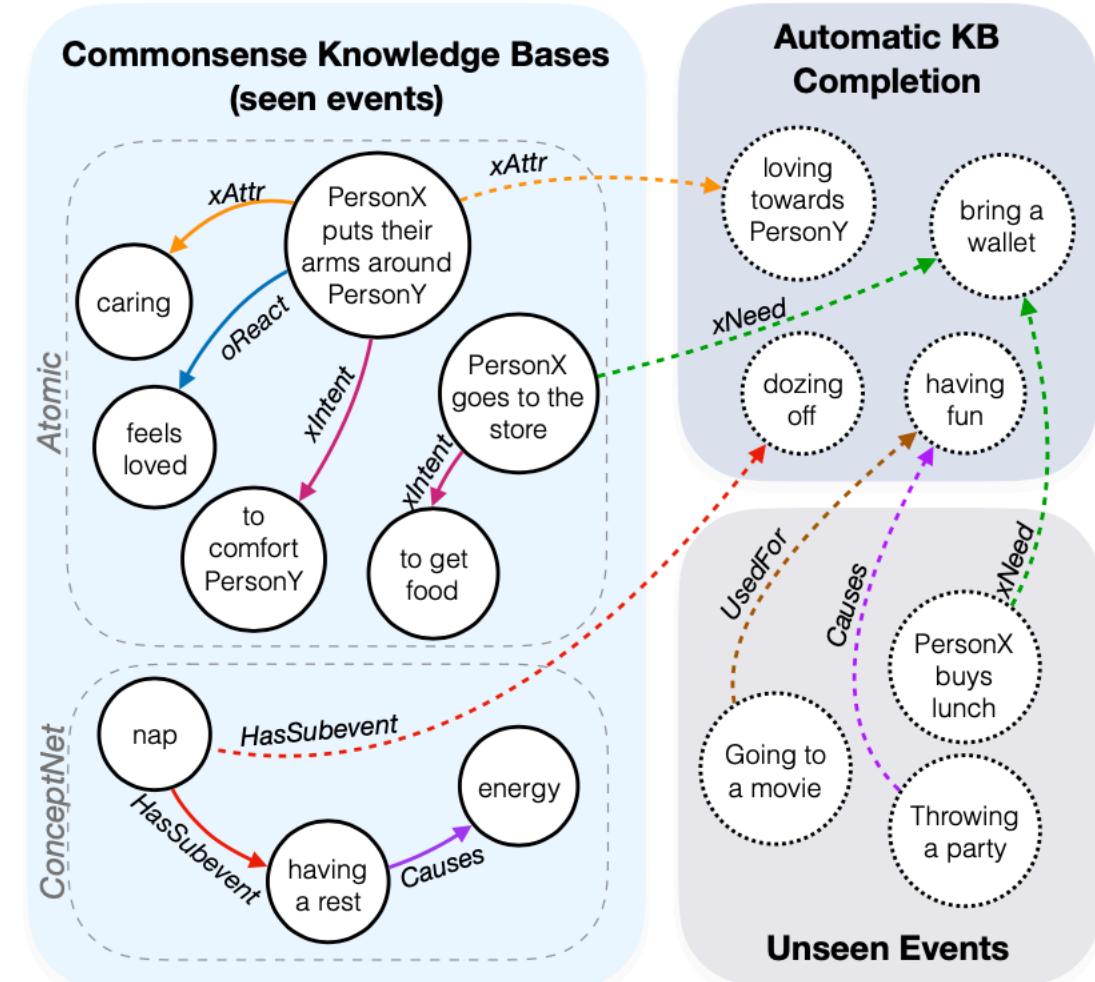
- The difficulty of achieving high concept coverage in high-precision curated KBs.
- In encyclopedic knowledge, the entities and relations are in a well-defined space. However, for commonsense knowledge, the relation between two entities can not fit into a schema.
- Beside knowledge mentioned in text, how to capture implicit commonsense knowledge (multi-hop knowledge) is also a challenge.

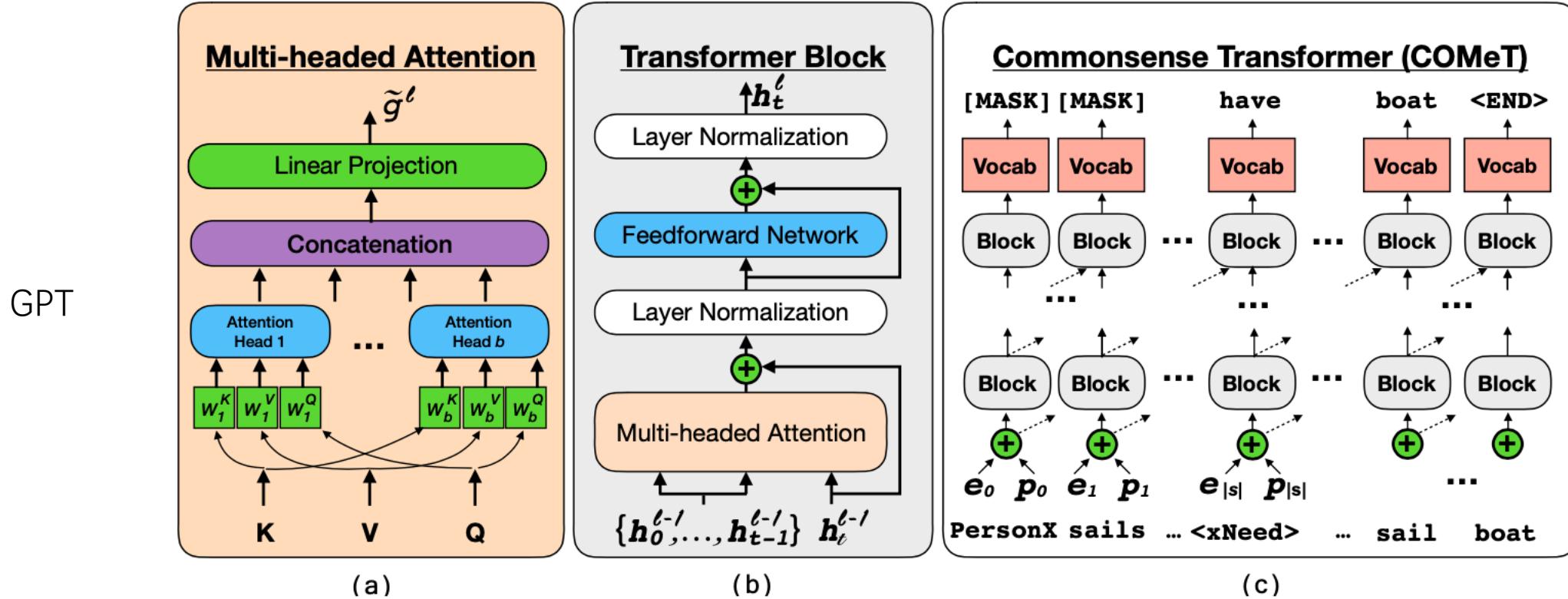
<https://mosaickg.apps.allenai.org/> (demo)

<https://github.com/atcbosselut/comet-commonsense> (code)

# Contribution

- A generative approach for knowledge base construction. (produce new nodes and identify edges between existing nodes).
- Large-scale transformer language models produce commonsense knowledge tuples by trained the seed tuples.
- Empirical study on the quality, novelty, and diversity of the commonsense knowledge produce for two domains, ATOMIC and ConceptNet.
- COMET is able to produce high quality tuples as human judges find that 77.5% of generated tuples for ATOMIC events and 91.7% of generated tuples for ConceptNet relations are correct.





Tuples: {s,r,o} format.

Given the concatenation of the tokens of s and r as input, the model must learn to generate the tokens of o.

#### ATOMIC Input Template and ConceptNet Relation-only Input Template



PersonX goes to the mall [MASK] <xIntent> to buy clothes

#### ConceptNet Relation to Language Input Template



go to mall [MASK] [MASK] has prerequisite [MASK] have money

# Experiments

<b>Model</b>	<b>PPL<sup>5</sup></b>	<b>BLEU-2</b>	<b>N/T <i>sro</i><sup>6</sup></b>	<b>N/T <i>o</i></b>	<b>N/U <i>o</i></b>
9ENC9DEC (Sap et al., 2019)	-	10.01	100.00	8.61	40.77
NearestNeighbor (Sap et al., 2019)	-	6.61	-	-	-
Event2(IN)VOLUN (Sap et al., 2019)	-	9.67	100.00	9.52	45.06
Event2PERSONX/Y (Sap et al., 2019)	-	9.24	100.00	8.22	41.66
Event2PRE/POST (Sap et al., 2019)	-	9.93	100.00	7.38	41.99
COMET (- pretrain)	15.42	13.88	100.00	7.25	45.71
COMET	<b>11.14</b>	<b>15.10</b>	100.00	<b>9.71</b>	<b>51.20</b>

Table 1: Automatic evalt scores are reported for the

<b>Model</b>	<b>PPL</b>	<b>Score</b>	<b>N/T <i>sro</i></b>	<b>N/T <i>o</i></b>	<b>Human</b>
LSTM - <i>s</i>	-	60.83	<b>86.25</b>	7.83	63.86
CKBG (Saito et al., 2018)	-	57.17	<b>86.25</b>	<b>8.67</b>	53.95
COMET (- pretrain)	8.05	89.25	36.17	6.00	83.49
COMET - RELTOK	4.39	95.17	56.42	2.62	<b>92.11</b>
COMET	<b>4.32</b>	<b>95.25</b>	59.25	3.75	91.69

Table 6: ConceptNet generation Results

monstence. No novelty in the training set.

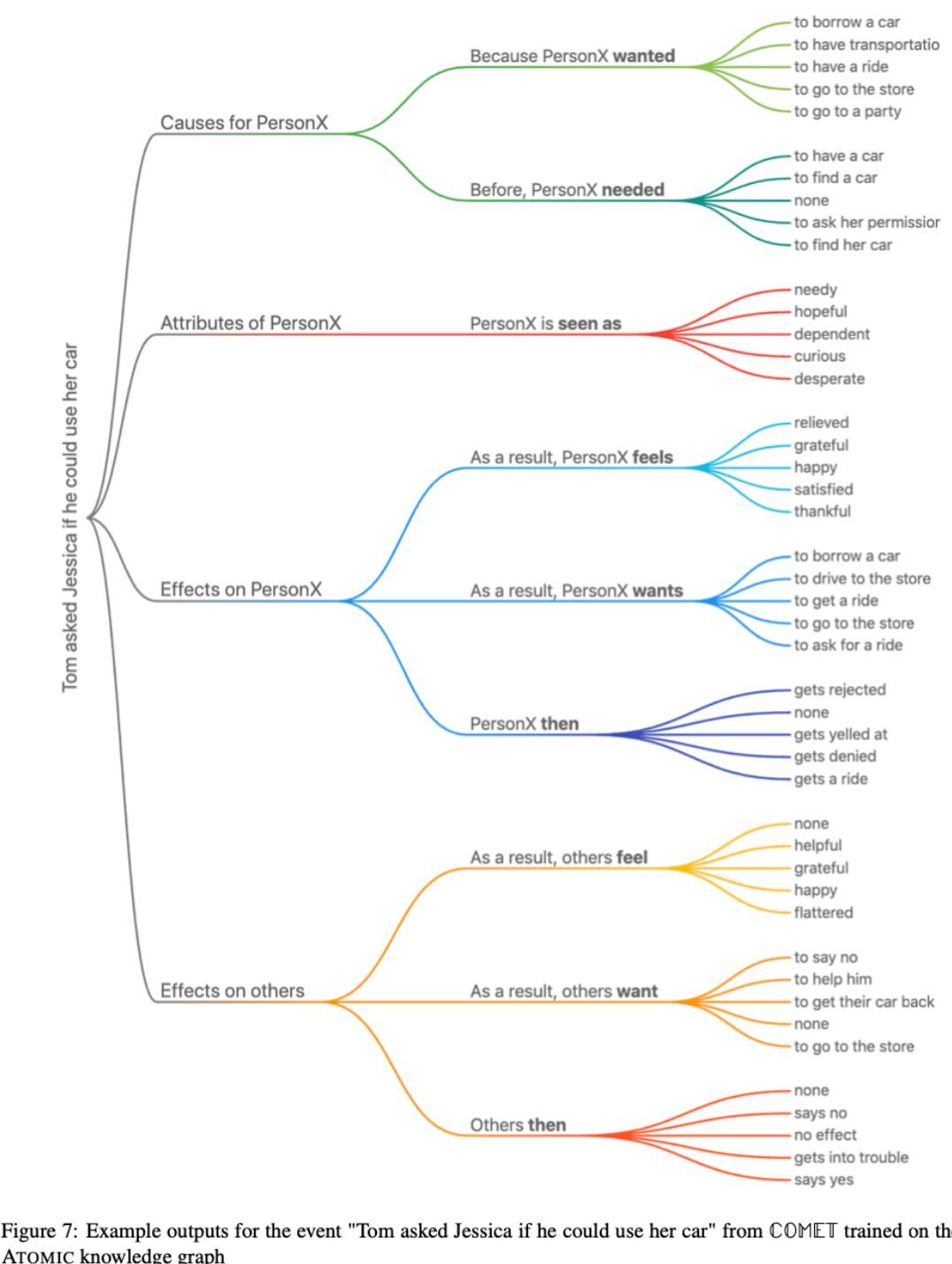
# Experiments

COMET Decoding method	<b>oEffect</b>	<b>oReact</b>	<b>oWant</b>	<b>xAttr</b>	<b>xEffect</b>	<b>xIntent</b>	<b>xNeed</b>	<b>xReact</b>	<b>xWant</b>	Avg
Top-5 random sampling (n=2500 per relation)	34.60	44.04	35.56	64.56	55.68	58.84	46.68	80.96	58.52	53.27
Top-10 random sampling (n=5000 per relation)	25.20	37.42	27.34	49.20	47.34	47.06	38.24	72.60	48.10	43.61
Beam search - 2 beams (n=1000 per relation)	43.70	54.20	47.60	<b>84.00</b>	51.10	73.80	50.70	85.80	78.70	63.29
Beam search - 5 beams (n=2500 per relation)	37.12	45.36	42.04	63.64	<b>61.76</b>	63.60	57.60	78.64	68.40	57.57
Beam search - 10 beams (n=5000 per relation)	29.02	37.68	44.48	57.48	55.50	68.32	64.24	76.18	75.16	56.45
Greedy decoding (n=500 per relation)	<b>61.20</b>	<b>69.80</b>	<b>80.00</b>	77.00	53.00	<b>89.60</b>	<b>85.60</b>	<b>92.20</b>	<b>89.40</b>	<b>77.53</b>
Human validation of gold ATOMIC	84.62	86.13	83.12	78.44	83.92	91.37	81.98	95.18	90.90	86.18

Table 3: Human evaluation testing effect of different decoding schemes on candidate tuple quality. The number of ratings made per relation for each decoding method is provided in the first column.

Seed Concept	Relation
X holds out X's hand to Y	xAttr
X meets Y eyes	xAttr
X watches Y every ____	xAttr
X eats red meat	xEffect
X makes crafts	xEffect
X turns X's phone	xEffect
X pours ____ over Y's head	oEffect
X takes Y's head off	oEffect
X pisses on Y's bonfire	oEffect
X spoils somebody rotten	xIntent
X gives Y some pills	xIntent
X provides for Y's needs	xIntent
X explains Y's reasons	xNeed
X fulfils X's needs	xNeed
X gives Y everything	xNeed
X eats pancakes	xReact
X makes ____ at work	xReact
X moves house	xReact
X gives birth to the Y	oReact
X gives Y's friend ____	oReact
X goes ____ with friends	oReact
X gets all the supplies	xWant
X murders Y's wife	xWant
X starts shopping	xWant
X develops Y theory	oWant
X offer Y a position	oWant
X takes ____ out for dinner	oWant

Table 5: Generations that come from a subset of **novel** generations in the development set. A novel generation is one not found in the training set. Margin indicates whether the tuple was annotated by a human annotator.



Completion	Plausible
machine	✓
food	✓
car	✓
feel	✓
subject	✓
fruit	✓
state	✓
space	✓
fridge	
your mind	🟡
town	✓
dentist	✓
your finger	
you feel good	✓
save life	✓
receive letter	✓
purity	✓
big	✓
fragile	✓
many plant	✓
play game	✓
ite	✓
ite	✓
ite	✓
ent	✓
die	✓
get wet	✓
bal	✓
you be tire	✓
on	✓
recycle	✓
wood	✓
planet	✓

ed and **novel** generations in the development set. Novel generations are not found in the training set. Margin indicates whether the tuple was annotated by a human annotator

# Summary

- COMET is a successful attempt for adapting the weights of language models to learn to produce commonsense knowledge tuples.
- Transformer is  and pre-training is  , too.
- Sparse issue of external knowledge. Can we stand on the original Kb and adapt kb construction?

# MuTual: A Dataset for Multi-Turn Dialogue Reasoning

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, Ming Zhou  
Zhejiang University

Microsoft Research Asia

School of Engineering, Westlake University

{cuileyang,zhangyue}@westlake.edu.cn

{Wu.Yu,shujliu,mingzhou}@microsoft.com

ACL 2020

# Background

There is still a huge gap between performance on the leader-board and user experience.

**MuTual**, a novel dataset for Multi-Reasoning, consisting of 8,860 man dialogues, based on Chinese student listening comprehension exams.

MuTual requires a model that can handle reasoning problems:

- Attitude Reasoning
- Algebraic Reasoning
- Intention Prediction
- Situational Reasoning
- Multi-fact Reasoning

Given a context, current systems are able to yield a relevant and fluent response, but sometimes make logical mistakes.

Context	Candidates Responses	Reasoning Type
<p>M: Hi, Della. How long are you going to stay here? F: Only 4 days. I have to go to London after the concert here at the weekend. M: I'm looking forward to that concert very much. Can you tell us where you <u>sing in public for the first time</u>? F: Hmm...at my <u>high school concert</u>, <u>my legs shook uncontrollably</u> and <u>I almost fell</u>.</p> <p>F: I'd like <u>2 tickets</u> for the 5:50 concert. M: That's <u>all be \$9</u>.</p>	<p>✓ M: Haha, I can imagine how nervous you were then. ✗ M: Why were you so nervous at that time? It wasn't your first singing at your high school concert. ✗ M: Yeah, if I had been you, I would have been happy too. ✗ M: Why did you feel disappointed?</p> <p>✗ F: Please give me \$9 refund. ✓ F: It's \$4.5 for each ticket, right? ✗ F: Shouldn't it be \$4.5 in total? ✗ F: I will pay you \$2 more.</p>	Attitude Reasoning (13%)
<p>F: I heard you were <u>having problems meeting your school fees</u> and <u>may not be able to study next term</u>. M: I was having some difficulties, but I have <u>received the scholarship</u> and <u>things are finally looking up</u>.</p>	<p>✗ F: Why are you going to drop out of school? ✗ F: You mean you'll try to get a scholarship? ✓ F: I am glad to hear that you will continue your studies. ✗ F: Why you have not received the scholarship?</p>	Intention Prediction (31%)
<p>F: Excuse me, sir. <u>This is a non smoking area</u>. M: Oh, sorry. I will move to the smoking area. F: I'm afraid <u>no table in the smoking area</u> is available now.</p>	<p>✗ M: Sorry. I won't smoke in the hospital again. ✓ M: OK. I won't smoke. Could you please give me a menu? ✗ M: Could you please tell the customer over there not to smoke? We can't stand the smell. ✗ M: Sorry. I will smoke when I get off the bus.</p>	Situation Reasoning (16%)
<p>M: This <u>painting</u> is one of the most valuable in the museum's collection. F: It is amazing. I'm glad I <u>spent \$30 on my ticket</u> to the exhibit today. M: <u>The museum purchased it in 1935 for \$2000</u>. But it is <u>now worth \$2,000,000</u>.</p>	<p>✗ M: I heard the museum purchased it in 1678 for \$2000. ✗ M: I heard the museum purchased it in 1678 for \$30. ✗ M: So the sculpture worth \$2,000,000 now. ✓ M: So the painting worth \$2,000,000 now.</p>	Multi-fact Reasoning (24%)
<p>M: Good evening, ma'am. Do you have a <u>reservation</u>? F: No, I don't. M: Awfully sorry, but there are <u>no empty tables left now</u>.</p>	<p>✓ F: The restaurant is too popular. ✗ F: The restaurant is not crowded at all. ✗ F: So I have to eat in a bad table in the restaurant. ✗ F: Show me the way to the table.</p>	Others (9%)

# Multi-turn dialogue datasets

dataset	Task	Reasoning	Domain	Manually
Ubuntu ( <a href="#">Lowe et al., 2015</a> )	<b>Next Utterances Prediction</b>	✗	Technique	✗
PERSONA-CHAT ( <a href="#">Zhang et al., 2018a</a> )	<b>Next Utterances Prediction</b>	✗	Persona	✓
Dialogue NLI ( <a href="#">Welleck et al., 2019</a> )	<b>Next Utterances Prediction</b>	✗	Persona	✗
CoQA ( <a href="#">Reddy et al., 2019</a> )	Conversational QA	✓	Diverse	✓
Douban ( <a href="#">Wu et al., 2017</a> )	<b>Next Utterances Prediction</b>	✗	Open	✗
DREAM ( <a href="#">Sun et al., 2019</a> )	Reading Comprehension	✓	Open	✓
WSC ( <a href="#">Levesque et al., 2012</a> )	Coreference Resolution	✓	Open	✗
SWAG ( <a href="#">Zellers et al., 2018</a> )	Plausible Inference	✓	Movie	✗
CommonsenseQA ( <a href="#">Talmor et al., 2019</a> )	Reading Comprehension	✓	Open	✓
RACE ( <a href="#">Lai et al., 2017</a> )	Reading Comprehension	✓	Open	✗
ARC ( <a href="#">Clark et al., 2018</a> )	Reading Comprehension	✓	Science	✗
DROP ( <a href="#">Dua et al., 2019</a> )	Reading Comprehension	✓	Open	✗
Cosmos ( <a href="#">Huang et al., 2019</a> )	Reading Comprehension	✓	Narrative	✓
MuTual	<b>Next Utterances Prediction</b>	✓	Open	✓

Table 1: Comparison between our dataset and other datasets. “Manually” indicates that human writing of the question or answers is involved in the data annotation process, rather than mere manual selection of data.

## Listening Comprehension

**Dialogue (Audio)**

M Ma'am, you forgot your phone.

F Oh, thanks, I couldn't live without this little thing.

M I know what you mean. It is of great significance to you. So did you enjoy your dinner? 这 topic 转移的，，，绝了

F Oh yes, everything was just perfect. It's so hard to take the whole family out to eat, but your restaurant was perfect. Johnny had his own place to play in and I had time to talk with my sisters and their husbands.

M I'm glad to hear it. Our kids area is always popular.

F Well, you can be sure we'll be back.

What is the probable relationship between the speakers?

✓ A. Waiter and Customer.  
✗ B. Brother and Sister.  
✗ C. Husband and Wife.

**Question & Answer**

**MuTual**

M Ma'am, you forgot your phone.

F Oh, thanks, I couldn't live without this little thing.

M I know what you mean. It is of great significance to you. So did you enjoy your dinner?

F Oh yes, everything was just perfect. It's so hard to take the whole family out to eat, but your restaurant was perfect. Johnny had his own place to play in and I had time to talk with my sisters and their husbands.

**Response**

✓ A. Thanks for your compliment for the restaurant. positive response

✗ B. I'm sorry that you don't have a good time. more negative response

✗ C. Goodbye brother! Love you. negative response

✗ D. Hurry up honey, or we will be late for the dinner. negative response

Figure 2: The process of modifying the listening comprehension test data.

Baseline category	Baseline method	Dev			Test		
		R@1	R@2	MRR	R@1	R@2	MRR
Baseline	Human Random	- 0.250	- 0.500	- 0.604	0.938 0.250	0.971 0.500	0.964 0.604
Individual scoring method (discrimination)	TF-IDF Dual LSTM (Lowe et al., 2015) SMN (Wu et al., 2017) DAM (Zhou et al., 2018) BERT (Devlin et al., 2019) RoBERTa (Liu et al., 2019)	0.276 0.266 0.274 0.239 0.657 <b>0.695</b>	0.541 0.528 0.524 0.463 0.867 0.878	0.541 0.538 0.575 0.575 0.803 0.824	0.279 0.260 0.299 0.241 0.648 <b>0.713</b>	0.536 0.491 0.585 0.465 0.847 <b>0.892</b>	0.542 0.743 0.595 0.518 0.795 <b>0.836</b>
Individual scoring method (generation)	GPT-2 (Radford et al., 2019) GPT-2-FT (Radford et al., 2019)	0.335 0.398	0.595 0.646	0.586 0.628	0.332 0.392	0.602 0.670	0.584 0.629
Multi-choice method	BERT-MC (Devlin et al., 2019) RoBERTa-MC (Liu et al., 2019)	0.661 0.693	0.871 <b>0.887</b>	0.806 <b>0.825</b>	0.667 0.686	0.878 0.887	0.810 0.822

Table 3: Comparison of varying approaches on MuTual.

Baseline category	Baseline method	Dev			Test		
		R@1	R@2	MRR	R@1	R@2	MRR
Baseline	Human Random	- 0.250	- 0.500	- 0.604	0.930 0.250	0.972 0.500	0.961 0.604
Individual scoring method (discrimination)	TF-IDF SMN (Wu et al., 2017) DAM (Zhou et al., 2018) BERT (Devlin et al., 2019) RoBERTa (Liu et al., 2019)	0.283 0.264 0.261 0.514 0.622	0.530 0.524 0.520 0.787 0.853	0.763 0.578 0.645 0.715 0.782	0.278 0.265 0.272 0.514 0.626	0.529 0.516 0.523 0.787 0.866	0.764 0.627 0.695 0.715 0.787
Individual scoring method (generation)	GPT-2 (Radford et al., 2019) GPT-2-FT (Radford et al., 2019)	0.305 0.226	0.565 0.577	0.562 0.528	0.316 0.226	0.574 0.611	0.568 0.535
Multi-choice method	BERT-MC (Devlin et al., 2019) RoBERTa-MC (Liu et al., 2019)	0.586 <b>0.621</b>	0.791 <b>0.830</b>	0.751 <b>0.778</b>	0.580 <b>0.643</b>	0.792 <b>0.845</b>	0.749 <b>0.792</b>
Transfer method	RoBERTa (Liu et al., 2019) RoBERTa-MC (Liu et al., 2019)	0.559 0.384	0.827 0.815	0.746 0.656	0.558 0.402	0.827 0.845	0.746 0.673

Table 4: Results on MuTual<sup>plus</sup>. Transfer method denotes that we train it on MuTual and test on MuTual<sup>plus</sup>.

F: Do you know what time it is right now in New York?  
M: Let me see. It's 5:00 pm now, in New York is 6 hours behind.

F: Let me see, 7 hours behind. It is 11:00 am now in New York.  
F: 5 hours ahead. It is 11:00 pm now in New York.

✗ F: Is it 5:00 pm as well?  
✓ F: It is 11:00 am now in New York.

F: Good morning. What can I do for you?  
M: I am looking for a flat for 2 people near the university.  
F: Well. There are several places available and the rent ranges from 80 to \$150 a month. What are your requirements?  
M: I think of flat for no more than \$100 a month is good. I prefer to live in a quiet street and I need at least 2 bedrooms.

✗ F: If you have any questions about enrollment, do not hesitate to ask me.  
✓ F: How about this flat? If you are satisfied, we can sign the contract tomorrow.

F: We have 2 floors in our supermarket.  
F: You want only 1 bedroom, so we have three flats that meet your requirement.

Figure 5: Error analysis. ✗ indicates RoBERTa-MC's prediction.

Instances that involve algebraic and situation show poor performance. These two reasoning types heavily depend on **commonsense reasoning**.

## Multi-turn understanding

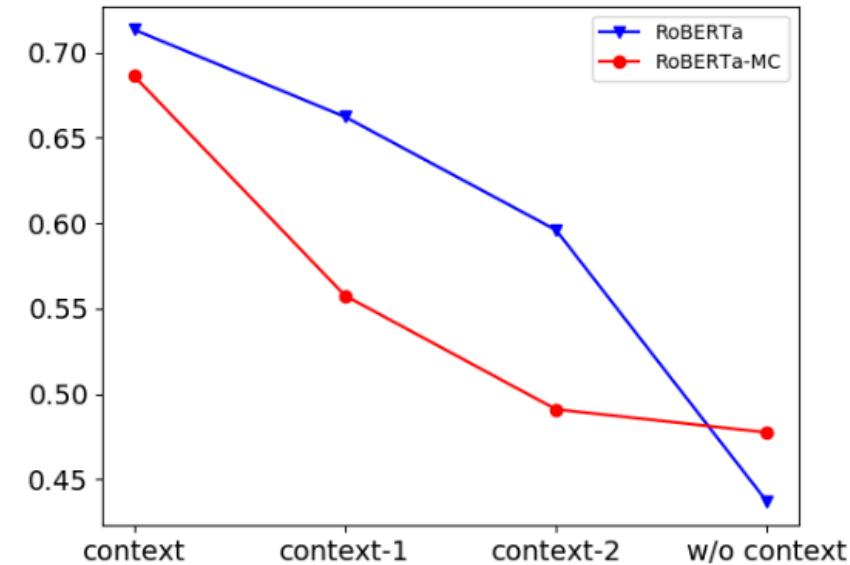


Figure 6: Ablation of context information. w/o context means all contexts are removed, so models just predict correct choice based on four candidates. context-n denotes the earliest n utterances are removed.

# Summary

- MuTual, a high-quality manually annotated multi-turn dialogue reasoning dataset, which contains 8,860 dialogues and aims to test reasoning ability of dialogue models.
- MuTual dataset is a next utterance prediction task, which is the fundamental problem in retrieval-based chatbots.
- Various state-of- the-art models show poor performance in MuTual. The best model RoBERTa only obtains 71.3% R@1.
- Moreover, if we shuffle the sequence of utterance, the performance of RoBERTa-MC drops by 3.8% only, showing that it is insensitive to the utterance sequence information (conversation flow).

# ASER: A Large-scale Eventuality Knowledge Graph

Hongming Zhang\*  
hzhagal@cse.ust.hk

Xin Liu\*  
xliucr@cse.ust.hk  
E, HKUST  
Hong Kong

Haojie Pan\*  
hpanad@cse.ust.hk  
CSE, HKUST  
Hong Kong

**Table 1: Size comparison of ASER and existing eventuality-related resources.** # Eventuality, # Relation, and # R types are the number of eventualities, relations between these eventualities, and relation types. For KGs containing knowledge about both entity and eventualities, we report the statistics about the eventualities subset. ASER (core) filters out eventualities that appear only once and thus has better accuracy while ASER (full) can cover more knowledge.

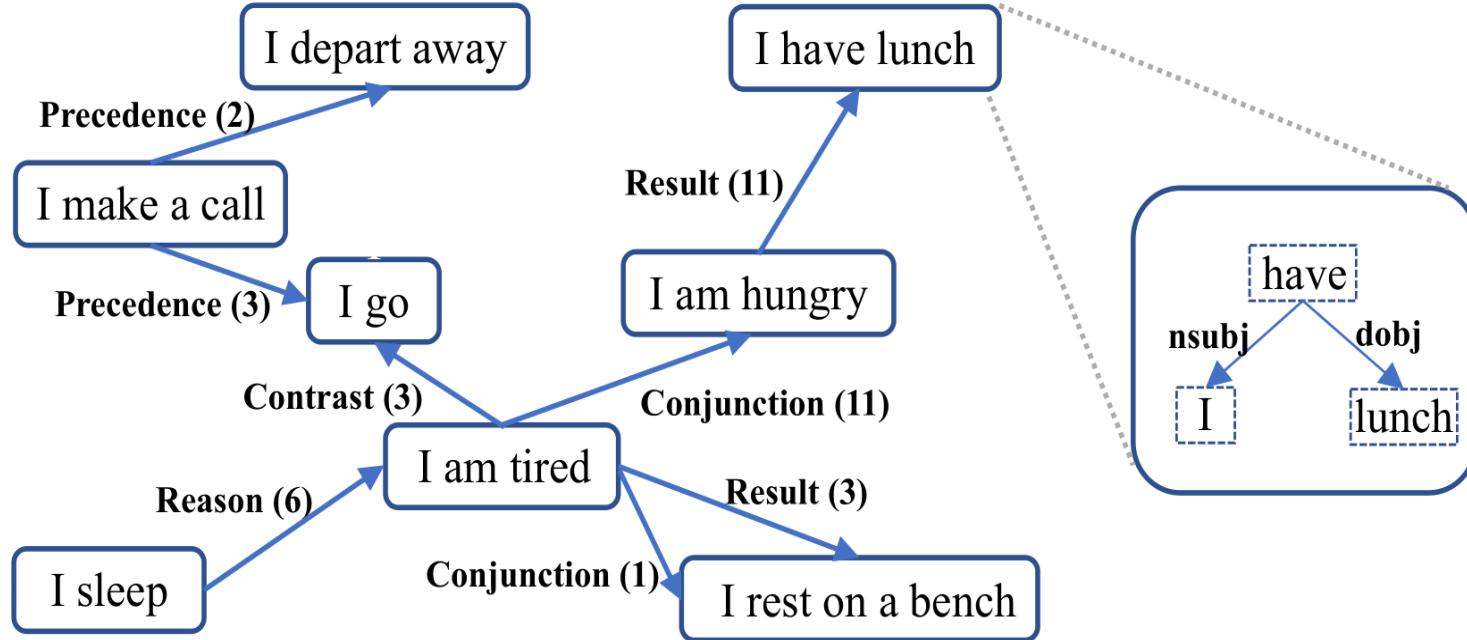
	# Eventuality	# Relation	# R Types
FrameNet [5]	27,691	1,709	7
ACE [2]	3,290	0	0
PropBank [30]	112,917	0	0
NomBank [26]	114,576	0	0
TimeBank [33]	7,571	8,242	1
ConceptNet [25]	74,989	116,097	4
Event2Mind [39]	24,716	57,097	3
ProPora [10]	2,406	16,269	1
ATOMIC [37]	309,515	877,108	9
Knowlywood [41]	964,758	2,644,415	4
ASER (core)	27,565,673	10,361,178	15
ASER (full)	194,000,677	64,351,959	15

Cane Wing-Ki Leung  
caneleung@wisers.com  
Wisers AI Lab  
Hong Kong

a large-scale **eventuality knowledge graph** extracted from the data.

Categories, 194-million unique eventualities, and 64-

large



Eventualities are connected with weighted directed edges. Each eventuality is a dependency graph.

ASER discovers useful real-world knowledge about Activities (or process, e.g., ‘I sleep’), States (e.g., ‘I am hungry’), Events (e.g., ‘I make a call’), and their Relations (e.g., ‘I am hungry’ may result in ‘I have lunch’), for which we call ASER.

ASER leverages carefully designed patterns to make sure the semantic completeness of extracted eventualities and uses a neural bootstrapping model to automatically learn relations between eventualities from large unlabeled corpus.

# Summary

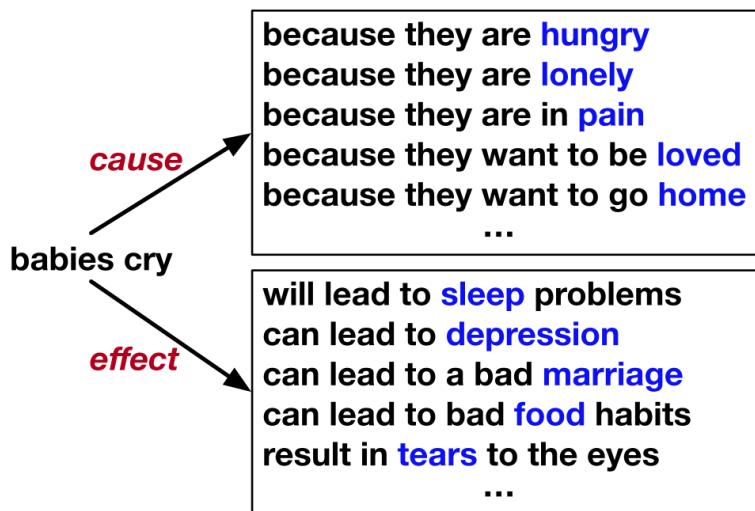
- ASER is a promising large-scale eventuality knowledge graph with great potential in many downstream tasks (QA, dialogue etc.).
- Inference over ASER is possible. Both eventuality and relation retrieval over one-hop or multi-hop relations can be modeled as conditional probability inference problems.
- The eventuality triples can be used to fine-tune the language model, which is shown to be very helpful.

# Guided Generation of Cause and Effect

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu and Benjamin Van Durme  
Harbin Institute of Technology, China  
Johns Hopkins University, USA {zyli,xding,tliu}@ir.hit.edu.cn,  
{edward.hu,vandurme}@jhu.edu

IJCAI 2020  
<http://openeg.8wss.com/generate/> (demo)

- proposing the task of open causal generation: producing possible causes and effects for any free-form textual event;
- construction of a causal corpus (CausalBank) containing 314 million CE (cause-effect) pairs; a large lexical causal knowledge graphs (Cause Effect Graph);



Sentential Causal Resource	# CE Pairs
TCR [Ning <i>et al.</i> , 2018]	172
SemEval-2007 Task4 [Girju <i>et al.</i> , 2007]	220
Causal-TimeBank [Mirza <i>et al.</i> , 2014]	318
CaTeRS [Mostafazadeh <i>et al.</i> , 2016]	488
EventCausalityData [Do <i>et al.</i> , 2011]	580
RED [O’Gorman <i>et al.</i> , 2016]	1,147
SemEval2010 Task8 [Hendrickx <i>et al.</i> , 2009]	1,331
BECAuSE 2.0 [Dunietz <i>et al.</i> , 2017b]	1,803
EventStoryLine [Caselli and Vossen, 2017]	5,519
PDTB 2.0 [Prasad <i>et al.</i> , 2008]	8,042
Altlex [Hidey and McKeown, 2016]	9,190
PDTB 3.0 [Webber <i>et al.</i> , 2019]	13 K
DisSent [Nie <i>et al.</i> , 2019]	167 K
<b>CausalBank (Ours)</b>	<b>314 M</b>
Causal Knowledge Graph	# CE Edges
Event2mind [Rashkin <i>et al.</i> , 2018]	25 K
ConceptNet 5.7 [Speer <i>et al.</i> , 2017]	473 K
ASER Core [Zhang <i>et al.</i> , 2019]	494 K
Atomic [Sap <i>et al.</i> , 2019]	877 K
CausalNet [Luo <i>et al.</i> , 2016]	13.3 M
<b>Cause Effect Graph (Ours)</b>	<b>89.1 M</b>

Table 5: Contrasting size with example prior works: only the causal portion of these corpora are listed. The top are sentential causal corpora, while the bottom are graph-structure causal knowledge bases.

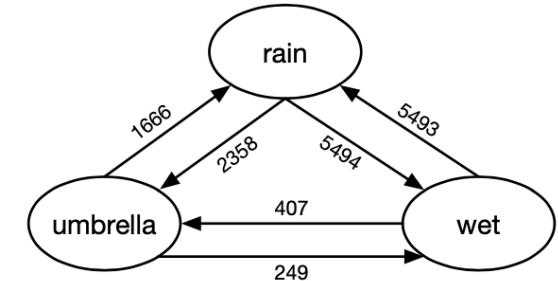


Figure 3: Cause Effect Graph: A lexical causal knowledge base.

输入

样例▼

## 包含

□

## 排除

□

## 生成目标

 结果  原因

## 语言

 中文  英文确认

## 输出

提示：点击删除按钮可替换低质量生成结果

交通事故	<span>删除</span>
刹车失灵	<span>删除</span>
发生侧翻	<span>删除</span>
连环追尾	<span>删除</span>
两车相撞	<span>删除</span>
打滑	<span>删除</span>

输入

样例▼

## 包含

□

## 排除

□

## 生成目标

 结果  原因

## 语言

 中文  英文确认

## 输出

提示：点击删除按钮可替换低质量生成结果

because she takes care of me	<span>删除</span>
because she is my mother	<span>删除</span>
because she makes me happy	<span>删除</span>
because she loves me unconditionally	<span>删除</span>

# Summary

- Retrieve causal knowledge and causal reasoning ability can play a support and explainable role for dialogue systems, reading comprehension systems, and so on.

# GPT-3

From JimmyLin SIGIR Talk

# 175 BILLION parameters!

I look at GPT-3 and I'm *not* depressed.

We know brute force works!

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples, often simple induction, something which current AI systems still largely struggle to do. Here we show that scaling up language models generally improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than previous non-autoregressive NLP models, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning.

I don't know, but the answer will be very exciting!

# GPT-3

Q: What is your favorite animal?

A: My favorite animal is a dog.

Q: Why?

A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?

A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?

A: A giraffe has two eyes.

Q: How many legs does a frog have?

A: A frog has four legs.

Q: Are there any animals with three legs?

A: No, there are no animals with three legs.

Q: Why don't animals have three legs?

A: Animals don't have three legs because they would fall over.

## Issues:

1. Logic problem with too long sequence.  
(Insufficient short-term memory)
2. Invalid questions that normal people will not ask or answer
3. Wrong question (wrong answer instead of refutation).

By Kevin Lacker

Thanks