DyKgChat: Benchmarking Dialogue Generation Grounding on Dynamic Knowledge Graphs

Yi-Lin Tuan Yun-Nung Chen Hung-yi Lee

National Taiwan University, Taipei, Taiwan
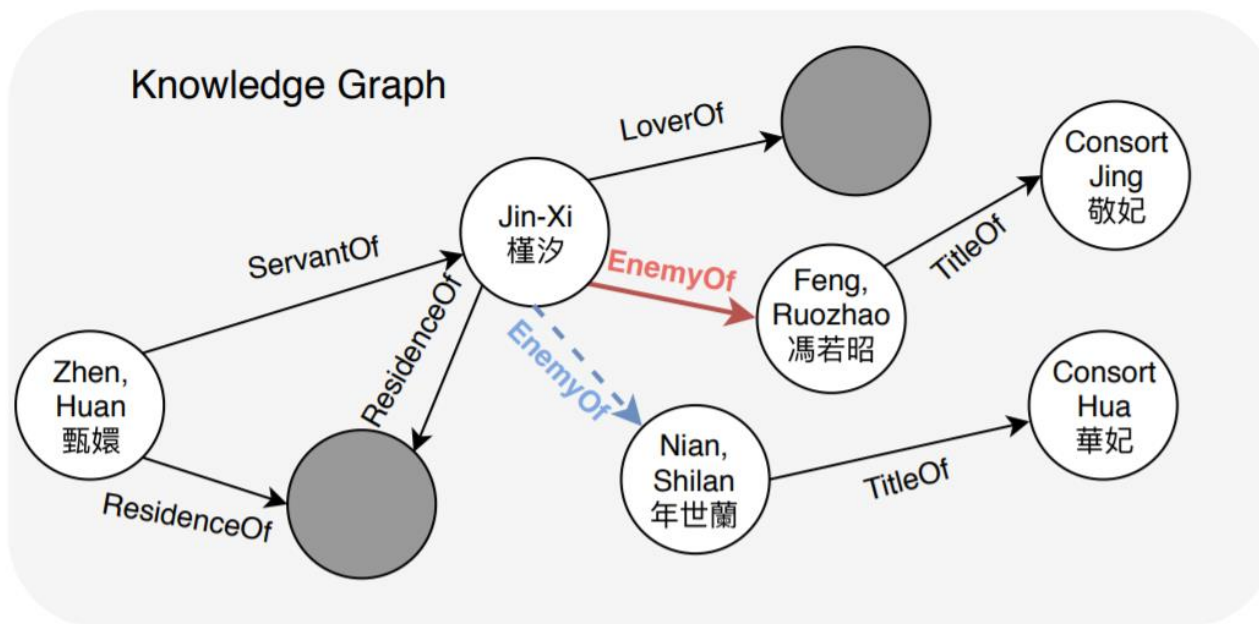
51205901082

陶明沺

Abstract

- New Task：
  - Apply dynamic knowledge graphs in neural conversation model
- Corpus:
  - A novel TV series conversation corpus (DyKgChat)
- Preliminary Model:
  - A sequence-to-sequencemodel (Seq2Seq)
  - A multi-hop reasoning model
- Benchmark:
  - Evaluation metrics and the experiments

Ideal Conversation Model

Task Description

- For each single-turn conversation:
  - Input message: x = $\{x_t\}_{t=1}^{m}$
  - Response: y = $\{y_t\}_{t=1}^{n}$
  - m and n are their lengths
  - Each turn (x, y) is paired with a knowledge graph K
  - K is composed of a collection of triplets (h, r, t)
  - h, t $\in$ V (the set of entities) and r $\in$ L (the set of relationships)

Task Description

- The task is two-fold:

  - Given an input message x and a knowledge graph K, the goal is to generate a sequence $\{\hat{y}_t\}_{t=1}^{n}$

  - After a knowledge graph is updated to K0 , the generated sequence should contain correspondent knowledge graph entities in K0 to reflect the updated information.

Evaluation Metrics

- Knowledge Entity Modeling
  - Knowledge word accuracy (KW-Acc)
  - Knowledge and generic word classification (KW/Generic)
  - Generated knowledge words (Generated-KW)
- Adaptation of Changed Knowledge Graphs
  - Change rate
  - Accurate change rate

DyKgChat Corpus

- Data Collection
  - "Hou Gong Zhen Huan Zhuang"
  - "Friends"

HGZHZ	**Zhen-Huan**: She must be frightened. It should blame me. I should not ask her to play chess. 甄嬛：姐姐定是嚇壞了。都怪臣妾不好，好端端的來叫梅姐姐下棋做什麼。 **Doctor-Wen**: Relax, **Concubine-Huan madame**. **Lady Shen** is just injured but is fine. 溫太醫：莞嬪娘娘請放心，惠貴人的精神倒是沒有大礙，只是傷口燒得有些厲害。
Friends	**Joey**: C' mon , you're going out with the guy! There's gotta be something wrong with him! **Chandler**: Alright **Joey**, be nice. So does he have a hump? A hump and a hairpiece?

Table 1: Examples of DyKgChat corpus.

DyKgChat Corpus

Metrics	HGZHZ	Friends
# Dialogues	1247	3092
Total # turns	17,164	57,757
Total # tokens	462,647	838,913
Avg. turns per dialogue	13.76	18.68
Avg. tokens per turn	26.95	14.52
Total unique tokens	3,624	19,762
# KG entities	174	281
# KG relation types	9	7
total # KG entities appear	46,059	176,550
# Dialogues w/ KG entities	1,166	2,373
# turns w/ KG entities	10,110	9,199

Table 2: The details of collected DyKgChat.

	Relation Type (Percentage)
HGZHZ	IsAliasOf (25%), IsChildOf (5%), IsLoverOf (6%), IsParentOf (5%), IsResidenceOf (16%), IsSiblingOf (2%), IsTitleOf (30%), IsEnemyOf (8%), IsServantOrMasterOf (3%)
Friends	IsLoverOf (12%), IsWorkplaceOf (2%), IsOccupationOf (8%), IsNameOf (47%), IsRelativeOf (8%), IsFriendOf (4%), IsNicknameOf (19%)

Table 3: The included relation types in the collect knowledge graphs, and their percentages.

DyKgChat Corpus

- Subgraph Sampling

  - A fixed knowledge graph G containing all information that once exists for each TV series

- Dynamic Knowledge Graphs:

  - The sample for each single-turn dialogue.

  - The sample for each multi-turn dialogue.

  - The manually-annotated subgraph for each period.

Qadpt

- Quick Adaptative Dynamic Knowledge-Grounded Neural Conversation Model

- The model is composed of

  - A Seq2Seq model with a controller

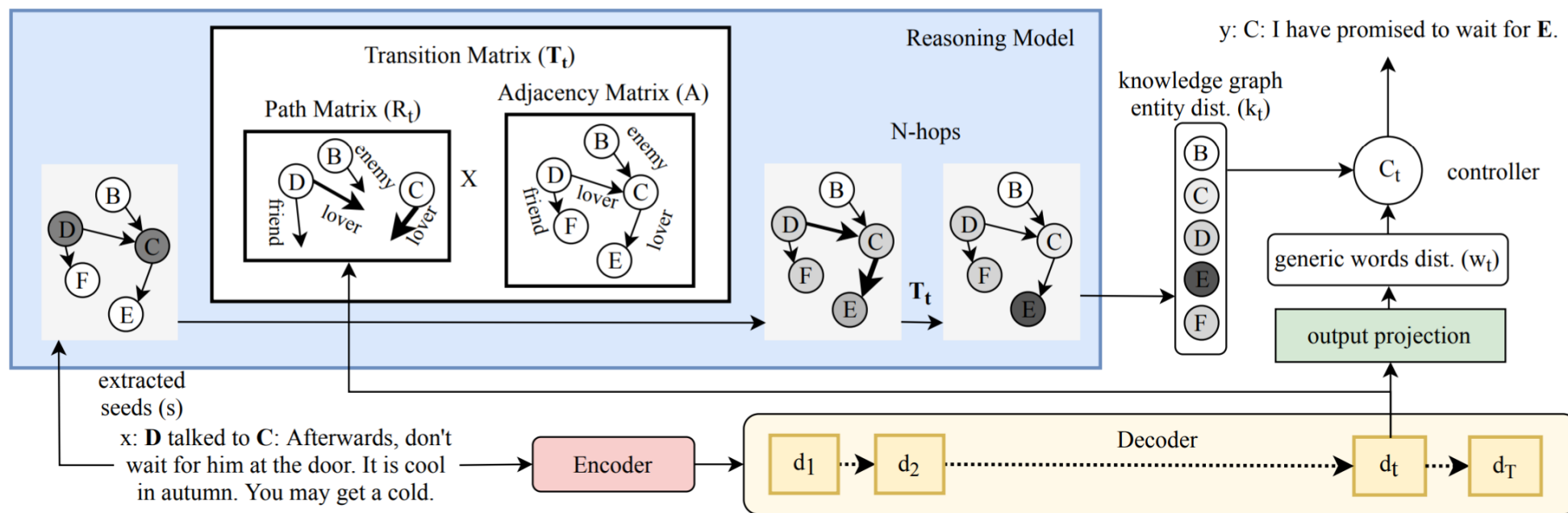 - A reasoning model

Qadpt

Figure 4: The framework of the proposed model. The node E here is the symbol for the emperor.

Experiments

- Baselines
    - Compare with prior knowledge-grounded conversation models:
        - The memory network and knowledgeaware model (KAware)
        - The topic-aware model (TAware)
        - MemNet

Experiments

- Results

Model	HGZHZ						Friends					
	Change Rate			Accurate Change Rate			Change Rate			Accurate Change Rate		
	All	Last1	Last2	All	Last1	Last2	All	Last1	Last2	All	Last1	Last2
MemNet	92.98	31.78	37.46	62.19	1.17	2.92	93.27	19.22	27.53	78.23	0.91	7.66
+ multi	98.69	77.87	81.96	**83.82**	3.40	10.74	95.31	28.09	36.65	**87.28**	0.69	7.63
TAware	94.38	68.33	71.86	**78.88**	1.95	9.26	92.93	26.52	30.96	**88.07**	0.35	10.63
+ multi	97.74	76.68	81.00	**95.30**	4.03	10.75	98.31	68.87	68.22	**92.29**	0.87	10.09
KAware	96.91	90.89	96.91	64.80	13.06	7.22	90.93	50.92	61.08	75.57	2.77	10.00
Qadpt	95.65	77.33	78.68	59.01	**66.67**	**16.82**	92.34	38.62	36.96	81.24	**30.85**	**16.87**
+ multi	99.60	83.17	87.27	56.11	**61.92**	**18.54**	98.47	48.78	63.54	86.97	**26.17**	**17.31**
+ TAware	99.02	83.14	85.59	58.82	**64.12**	**14.90**	98.45	56.77	65.25	82.52	**28.34**	**17.68**

Table 4: The results of change rate and accurate change rate.

Model	HGZHZ					Friends				
	KW	KW/Generic		Generated-KW		KW	KW/Generic		Generated-KW	
	Acc	Recall	Precision	Recall	Precision	Acc	Recall	Precision	Recall	Precision
Seq2Seq	12.10	29.08	27.44	13.30	24.28	3.81	23.22	5.57	6.88	2.02
MemNet	22.58	39.09	33.41	39.52	67.10	22.79	37.18	24.58	46.02	53.98
+ multi	35.20	54.49	30.15	60.63	83.43	34.92	47.31	22.78	60.54	69.46
TAware	50.21	44.40	35.50	49.18	76.72	62.74	**50.78**	22.50	57.84	62.83
+ multi	**59.71**	**68.61**	28.70	**70.18**	**85.54**	**72.96**	42.98	25.74	**71.11**	**77.35**
KAware	20.53	40.63	**36.64**	24.61	43.13	13.52	30.76	**28.42**	15.14	18.74
Qadpt	**57.61**	38.24	28.31	44.50	**90.70**	**74.00**	41.33	25.31	69.30	**77.30**
+ multi	**57.40**	51.97	28.43	**64.55**	**91.22**	**74.44**	42.81	25.01	**74.63**	**77.09**
+ TAware	**56.24**	53.68	31.03	**63.66**	**88.99**	**73.57**	47.05	25.91	**74.52**	**78.56**

Table 5: The results of knowledge graph entities prediction.

Experiments

- Human Evaluation

Model	HGZHZ						Friends					
	BLEU	PPL	dist-1	dist-2	dist-3	dist-4	BLEU	PPL	dist-1	dist-2	dist-3	dist-4
Seq2Seq	14.20	94.48	0.008	0.039	0.092	0.150	15.46	73.23	0.004	0.016	0.026	0.032
MemNet	**15.73**	88.29	0.012	0.062	0.150	0.240	14.61	67.58	0.005	**0.023**	0.040	0.049
+ multi	15.88	86.76	0.010	0.058	0.138	0.224	12.97	**54.67**	**0.006**	0.022	0.032	0.036
TAware	15.07	81.54	0.013	0.068	0.153	0.223	14.78	60.61	0.002	0.007	0.013	0.016
+ multi	13.34	**80.48**	**0.022**	0.122	0.239	0.304	15.74	56.67	0.003	0.011	0.019	0.023
KAware	14.14	90.11	0.011	0.061	0.135	0.198	15.70	64.70	0.002	0.009	0.017	0.021
Qadpt	14.52	88.24	0.013	0.081	0.169	0.242	**17.01**	68.27	0.002	0.008	0.013	0.016
+ multi	15.47	86.65	0.021	**0.129**	**0.259**	**0.342**	14.79	66.70	0.005	**0.023**	**0.041**	**0.051**
+ TAware	15.05	81.75	0.022	0.123	0.246	0.332	16.85	55.46	0.003	0.012	0.020	0.024

Table 6: The results of responses generation with BLEU, perplexity (PPL), distinct scores (1-gram to 4-gram).

Conclusion

- New task：
    - Dynamic knowledge-grounded conversation generation
- New dataset:
    - DyKgChat
- Evaluation metrics and baseline models:
    - Motivate the future research directions.

THANKS