Knowledge and Conversation Systems

李俊涛

2019.4.19

Paper List

- Towards Exploiting Background Knowledge for Building Conversation Systems, EMNLP 18.

- Wizard of Wikipedia: Knowledge-Powered Conversational Agents, ICLR 19.

Why these two papers?

- Knowledge is a research hotspot for dialogue and text generation
- Both two papers are proposed for constructing new task and new dataset (other than knowledge graph).
- Open conversation for specific-domain/open-domain.

Background and Motivation

- Treating conversation as a sequence to sequence generation task is an overly simplistic view V.S. humans converse by heavily relying on their background knowledge about a topic).

- Previous work attempt to incorporate external knowledge with conversation generation V.S. utterances are explicitly linked to external knowledge.

Plot

... The lab works on spiders and has even managed to create new species of spiders through genetic manipulation. While Peter is taking photographs of Mary Jane for the school newspaper, one of these new spiders lands on his hand and bites him Peter comes home feeling ill and immediately goes to bed. ...

Review

... I thoroughly enjoyed "Spider-Man" which I saw in a screening. I thought the movie was very engrossing. Director Sam Raimi kept the action quotient high, but also emphasized the human element of the story. Tobey was brilliant as a gawky teenager...

Movie: Spider-Man

Speaker 1(N): Which is your favourite character?

Speaker 2(C): My favorite character was Tobey Maguire.

Speaker 1(N): I thought he did an excellent job as peter parker, I didn't see what it was that turned him into Spider-Man though.

Speaker 2(P): Well this happens while Peter is taking photographs of Mary Jane for the school newspaper, one of these new spiders lands on his hand and bites him.

Speaker 1 (N): I see. I was very excited to see this film and it did not disappoint!

Speaker 2(R): I agree, I thoroughly enjoyed "Spider-Man"

Speaker 1(N): I loved that they stayed true to the comic.

Speaker 2(C): Yeah, it was a really great comic book adaptation

Speaker 1(N): The movie is a great life lesson on balancing power.

Speaker 2(F): That is my most favorite line in the movie, "With great power comes great responsibility."

Comments

... Crazy attention to detail. My favorite character was Tobey Maguire. I can't get over the "I'm gonna kill you dead" line. It was too heavily reliant on constant light-hearted humor. However the constant joking around kinda bogged it down for me. A really great comic book adaptation. ....

Fact Table

Awards	Golden Trailer Awards 2002
Taglines	With great power comes great responsibility. Get Ready For Spidey !
Similar Movies	Iron Man Spider-Man 2

Dataset Construction

- Curating a list of popular movies
- Collecting background knowledge
- Meta data or Fact Table
- Collecting conversation starters
- Collecting background knowledge aware conversations via crowdsourcing
- Verification of the collected chats

Collecting background knowledge

- Review(R): fetch the top 2 most popular reviews for each movie with more than 50 words.
- Plot(P): extract information about the plot of each movie from the correlated Wikipedia page.
- Comments(C): Reddit, comments about various aspects of movie.
- Fact Table (F): 4 fields, box office collection, similar movies, awards, tag-lines, and the reason for using 4 fields.

Collecting background knowledge aware conversations

- Conversation starters

Observation: workers converse with a lot of the initial turns in greetings and general chit-chat before actually chatting about a movie.

Solutions: three exactly opening statements as conversation starters (favorite scene, favorite character, opinion).

- Conversations

Given the background knowledge and conversation starters, each worker is asked to create at least 8 utterances using the self-chat strategies.

Verification of the collected chats

- Every chat that was collected by the above process was verified by an in-house evaluator to check if the workers adhered to the instructions and produced coherent chats.
- Verify whether these conversations are natural and rate them on five different parameters:

  Intelligible, i.e., an average reader could understand the conversation

  Coherent, there were no abrupt context switches

  Grammatically correct

  On topic, the chat revolved around the concerned movie limited to related movies/characters/actors

  Natural two-person chats, i.e., the roleplay setup does not make the chat look unnatural

Metric	Rating	α	κ
Intelligible	4.47 ± 0.52	0.70	0.69
Coherent	4.33 ± 0.93	0.57	0.71
Grammar	4.41 ± 0.56	0.60	0.69
Two-person-chat	4.47 ± 0.46	0.64	0.70
On Topic	4.57 ± 0.43	0.72	0.70

Table 1: Average human evaluation scores with standard deviations for conversations (scale 1-5). We also report mean Krippendorff's α and mean Cohen's κ

#chats	9071
#movies	921
#utterances	90810
Average # of utterances per chat	10.01
Average # of words per utterance	15.29
Average # of words per chat	153.07
Average # of words in Plot	186.10
Average # of words in Review	384.44
Average # of words in Comments	123.81
Average # of words in Fact Table	33.47
# unique Plots	5157
# unique Reviews	1817
# unique Comments	12740

Table 2: Statistics of the dataset

Models

- Generation-based Model: HERD
- Generate-or-copy models: Pointer Network
- Span prediction models: Bi-directional attention flow model (BiDAF),

Given a document and a question, the model predicts the span in the document which contains the answer.

Experimental Setup

- Creating train/valid/test splits
- Creating training instances
- Merging resources into a single document
- Evaluation metrics: automatic evaluation and human evaluations (generation and span prediction)
- Collecting multiple reference responses

Results

Model	F1		BLEU		Rouge-1		Rouge-2		Rouge-L	
HRED	-	-	5.23	5.38	24.55	25.38	7.61	8.35	18.87	19.67
GTTP (o)	-	-	13.92	16.46	30.32	31.6	17.78	21.21	25.67	27.83
GTTP (ms)	-	-	11.05	15.68	29.66	31.71	17.70	19.72	25.13	27.35
GTTP (ml)	-	-	7.51	8.73	23.20	21.55	9.91	10.42	17.35	18.12
BiDAF (o)	39.69	47.18	28.85	34.98	39.68	46.49	33.72	40.58	35.91	42.64
BiDAF (ms)	45.73	51.35	32.95	39.39	45.69	50.73	40.18	45.01	43.46	46.95

Table 3: Performance of the proposed models on our dataset. The figures on the left in each column indicate scores on single-reference test dataset while the figures on the right denote scores on multi-reference dataset.

Background and Motivation

- to be able to comprehend language,
- employ memory to retain and recall knowledge,
- to reason about these concepts together
- output a response that both fulfills functional goals in the conversation while simultaneously being captivating to their human speaking partner.

Task Formulation

- two participants engage in chitchat, with one of the participants selecting a beginning topic, and during the conversation the topic is allowed to naturally change.
- The two participants, however, are not quite symmetric: one will play the role of a knowledgeable expert (which we refer to as the wizard) while the other is a curious learner (the apprentice).

Conversation Flow

- Either the wizard or apprentice is picked to choose the topic and speak first. The other player receives the topic information, and the conversation begins.
- When the apprentice sends the wizard a message, the wizard is shown relevant knowledge (described below), and chooses a relevant sentence in order to construct a response, or else chooses the no sentence used option.
- The Wizard responds to the apprentice basing their response on their chosen sentence.
- The conversation repeats until one of the conversation partners ends the chat (after a minimum of 4 or 5 turns each, randomly chosen beforehand).

Statistics

Table 1: Dataset statistics of the Wizard of Wikipedia task.

Wizard of Wikipedia Task	Train	Valid	Test Seen	Test Unseen
Number of Utterances	166,787	17,715	8,715	8,782
Number of Dialogues	18,430	1,948	965	968
Number of Topics	1,247	599	533	58
Average Turns per Dialogue	9.0	9.1	9.0	9.1
Knowledge Database	5.4M articles		93M sentences	

Model

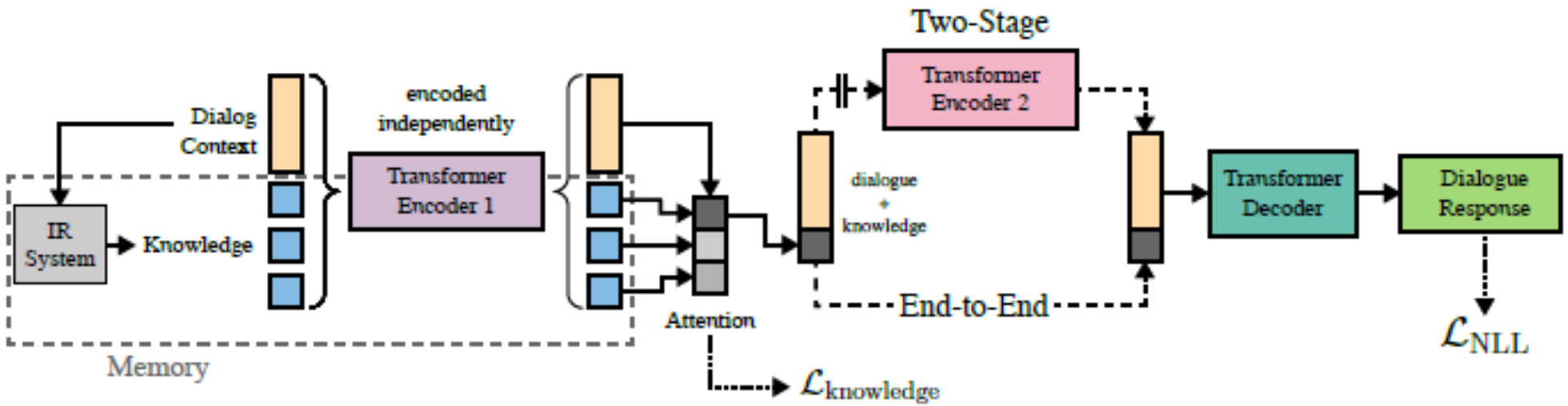

Figure 1: **Generative Transformer Memory Network.** An IR system provides knowledge candidates from Wikipedia. Dialogue Context and Knowledge are encoded using a shared encoder. In the Two-stage model, the dialogue and knowledge are re-encoded after knowledge selection.

Table 2: **Test performance of various methods on the Knowledge Selection Task.** The models must select the gold knowledge sentences chosen by humans given the dialogue context.

	Seen Test		Unseen Test	
Method	R@1	F1	R@1	F1
Random	2.7	13.5	2.3	13.1
IR baseline	5.8	21.8	7.6	23.5
BoW MemNet	23.0	36.3	8.9	22.9
Transformer	22.5	33.2	12.2	19.8
Transformer (+Reddit pretraining)	24.5	**36.4**	**23.7**	**35.8**
Transformer (+Reddit pretraining, +SQuAD training)	**25.5**	36.2	22.9	34.2

Table 3: **Retrieval methods on the full Wizard task.** Models must select relevant knowledge and retrieve a response from the training set as a dialogue response. Using knowledge always helps, and the Transformer Memory Network with pretraining performs best.

	Predicted Knowledge				Gold Knowledge	
	Test Seen		Test Unseen		Seen	Unseen
Method	R@1	F1	R@1	F1	R@1	R@1
Random	1.0	7.4	1.0	7.3	1.0	1.0
IR baseline	17.8	12.7	14.2	11.6	73.5	67.5
BoW MemNet (no knowledge)	56.1	14.2	28.8	11.6	56.1	28.8
BoW MemNet	71.3	**15.6**	33.1	12.3	84.5	66.7
Transformer (no knowledge, w/o Reddit)	60.8	13.3	25.5	9.7	60.8	25.5
Transformer (no knowledge, w/ Reddit)	79.0	15.0	54.0	11.6	79.0	54.0
Transformer MemNet (w/ Reddit)	86.8	15.4	**69.8**	**12.4**	91.6	82.3
Transformer MemNet (w/ Reddit+SQuAD)	**87.4**	15.4	**69.8**	**12.4**	**92.3**	**83.1**

Table 4: **Generative models on the full Wizard Task.** The Two-stage model performs best using predicted knowledge, while the End-to-end (E2E) model performs best with gold knowledge.

	Predicted Knowledge				Gold Knowledge			
	Test Seen		Test Unseen		Test Seen		Test Unseen	
Method	PPL	F1	PPL	F1	PPL	F1	PPL	F1
Repeat last utterance	-	13.8	-	13.7	-	13.8	-	13.7
Transformer (no knowledge)	-	-	-	-	41.8	17.8	87.0	14.0
E2E Transformer MemNet (no auxiliary loss)	66.5	15.9	103.6	14.3	24.2	33.6	35.5	29.5
E2E Transformer MemNet (w/ auxiliary loss)	63.5	16.9	97.3	14.4	**23.1**	**35.5**	**32.8**	**32.2**
Two-Stage Transformer MemNet	54.8	18.6	88.5	**17.4**	30.0	30.7	42.7	28.6
Two-Stage Transformer MemNet (w/ K.D.)	**46.5**	**18.9**	**84.8**	17.3	28.6	30.6	43.7	28.0

Thank You !