Appendix for Towards Exploiting Background Knowledge for Building Conversation Systems

Model details - GTTP

Since we modified the existing architecture of Get to the Point (See et al., 2017), we now provide details of the same. In the summarization task, the input is a *document* and the output is a *summary* whereas in our case the input is a {*resource/document, context*} pair and the output is a *response*. Note that the context includes the previous two utterances (dialog history) and the current utterance. Since, in both the tasks, the output is a sequence (*summary* v/s *response*) we don't need to change the decoder (*i.e.*, we can use the decoder from the original model as it is). However, we need to change the input fed to the decoder. Similar to the original model, we use an RNN to compute the representation of the document. Let N be the length of the document then the RNN computes representations $h^r_1, h^r_2, ..., h^r_N$ for all the words in the resource (we use the superscript r to indicate resource). The final representation of the resource is then the attention weighted sum of these word representations:

$$\begin{aligned} e^t_i &= v^T tanh(W_r h^r_i + U s_t + b_r) \\ a^t &= softmax(e^t) \\ r_t &= \sum_i a^t_i h^r_i \end{aligned} \tag{1}$$

where s_t is the state of the decoder at the current time step. In addition, in our case, we also have the context of the conversation apart from the document (resource). Once again, we use an RNN to compute a representation of this context. Specifically, we consider the previous k utterances as a single sequence of words and feed these to an RNN. Let M be the total length of the context (*i.e.*, all the k utterances taken together) then the RNN computes representations $h^c_1, h^c_2, ..., h^c_M$ for all the words in the context (we use the superscript c to indicate context). The final representation of the context is then the attention weighted sum of these word representations:

$$\begin{aligned} f^t_i &= v^T tanh(W_c h^c_i + V s_t + b_c) \\ m^t &= softmax(f^t) \\ c_t &= \sum_i m^t_i h^c_i \end{aligned} \tag{2}$$

where s_t is the state of the decoder at the current time step.

The decoder then uses r_t (document representation), c_t (context representation) and s_t (decoder's internal state) to compute a probability distribution over the vocabulary P_{vocab}. In addition the model also computes p_{gen} which indicates that there is a probability p_{gen} that the next word will be *generated* and a probability $(1 - p_{gen})$ that the next word will be *copied*. We use the following modified equation to compute p_{gen}

$$p_{gen} = \sigma(w^T_r r_t + w^T_c c_t + w^T_s s_t + w^T_x x_t + b_g) \tag{3}$$

where x_t is the previous word predicted by the decoder and fed as input to the decoder at the current time step. Similarly, s_t is the current state of the decoder computed using this input x_t. The final probability of a word w is then computed using a combination of two distributions, *viz.*, (P_{vocab}) as described above and the attention weights assigned to the document words as shown below

$$P(w) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i = w} a^t_i \tag{4}$$

where a^t_i are the attention weights assigned to every word in the document as computed in Equation 1. Thus, effectively, the model could learn to copy a word i if p_{gen} is low and a^t_i is high.

Example from the multiple reference test set

As seen from the Table 1, the given chat on "Secret Life of Pets" can have multiple responses for Speaker 2. Notice how Reference 1 talks against low critique scores thus emphasizing that he was totally impressed by the movie while Reference 4 has neutral opinion about the same. At the same time, Reference 3 talks about movie specific details like his favorite character while Reference 4 gives a personal opinion. All these four responses are valid given the current context.

Hyper-parameters

We describe the hyperparameters that we used for each model in this sub section. Following the original paper, we trained the HRED model using Adam (Kingma and Ba, 2014) optimizer with an initial learning rate of 0.001 on a minibatch of size 16. We used a dropout (Srivastava et al., 2014) with a rate of 0.25. For word embeddings we use pre-trained GloVe (Pennington et al., 2014) embeddings of size 300. For all the encoders and decoders in the model we used Gated Recurrent Unit (GRU) with 300 as the size of the hidden state. We restricted our vocabulary size to 20,000 most frequent words.

We followed the hyperparameters mentioned in the original paper and trained GTTP using Adagrad (Duchi et al., 2011) optimizer with an initial learning rate of 0.15 and an initial accumulator value of 0.1 on a minibatch of size 16. For the encoders and decoders we used LSTMs with 256 as the size of the hidden state. To avoid vanishing and exploding gradient problem we use gradient clipping with a maximum gradient norm of 2. We used early stopping based on the validation loss.

Again following the original paper, we trained BiDAF using AdaDelta (Zeiler, 2012) optimizer with an initial learning rate of 0.5 on a minibatch of size 32. For all encoders, we use LSTMs with 256 as the size of the hidden state. We used a dropout (Srivastava et al., 2014) rate of 0.2 across all LSTM layers, and for the linear transformation before the softmax for the answers. For word embeddings we use pre-trained GloVe embeddings of size 100. For both GTTP and BiDAF, we had to restrict context length to 65 tokens for fair comparison. Note that GTTP can scale beyond 65 tokens but BiDAF cannot.

Sample responses produced by the models

As seen from Table 2, HRED isn't able to produce responses that correspond to the given movie or the given context as it lacks any notion of background knowledge associated with it. We will not consider HRED for the following discussion. In Example 1, we can clearly see that only GTTP (oracle) matches with the ground truth. The remaining three models produce varied outputs which are still relevant to the context. In Example 2, we observe that prediction based models produce appropriate recommendation because of better context-document mapping mechanisms. Both the GTTP models produce responses which are copied but irrelevant to the context. At the same time, just producing spans without any structure isn't natural. This explains the need for hybrid models. Example 3 asks for the backstory of a character which requires complex reasoning. The model has to first understand the plot of the movie to locate the sentences which talk about that character's past. As seen from the responses of the given models, all of them except GTTP (ml) pick sentences which are relevant to the character but do not answer the required question. As discussed earlier, these models rely on word-overlap and thus possess limited natural language understanding. Thus, we need models which are capable of going beyond word overlap and producing responses even in such complex scenarios.

1 Collection of popular movies list

Following are the urls used to curate the popular movies list:
IMDb top 250 :
https://www.imdb.com/chart/top
Popular movies by genre:
https://www.imdb.com/feature/genre/
Other popular movies lists:
http://www.filmsite.org/guinness.html
https://www.thetoptens.com/best-movie-genres/
http://1001films.wikia.com/wiki/The_List
The imdb ids of the movies will be made available with every example.

Movie Name	The Secret Life of Pets
Chat	Speaker1 : What do you think about the movie? Speaker2 : I think it was comical and entertaining. Speaker1 : It delivered what was promised.
Reference 1	I agree! I'm surprised this film got such a low overall score by users.
Reference 2	My favorite character was Gidget! She was so much fun and so loyal to her friends!
Reference 3	Yes! As a Great Dane owner, I often wonder what my dogs are thinking. It was fun to see this take on it.
Reference 4	It was full of cliches with a predictable story, but with some really funny moments.

Table 1: Multiple references for the given chat.

Data Collection Interfaces used on Amazon Mechanical Turk

As explained in the paper, we resorted to AMT for three types of data collection *viz.* collection of opening statements as discussed in stage 3 of the dataset collection (Figure 5), the actual chat data collection explained in stage 4 of the procedure and additional responses collected for the test dataset. We will now show the instruction screens. The html files for the same will be released along with the code.

References

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

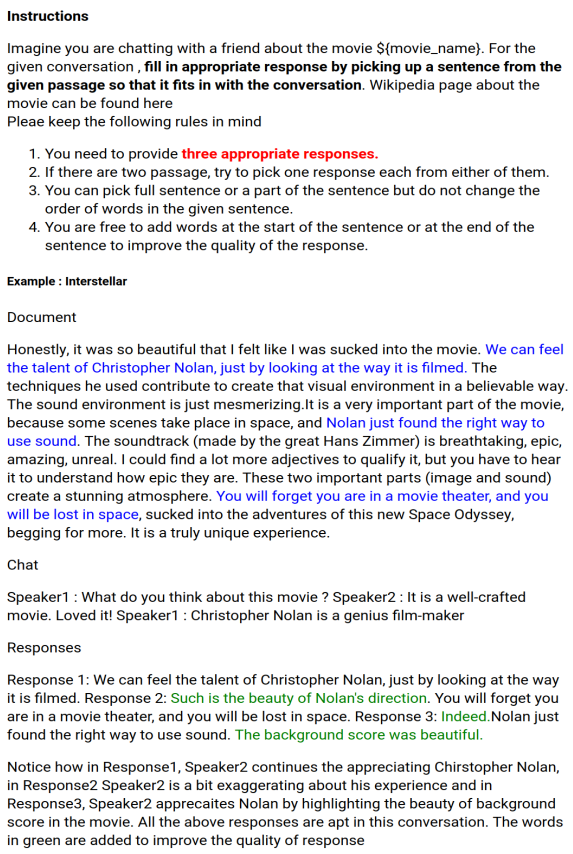
Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Movie	Example 1 (Creed)	Example 2 (Ted)	Example 3 (Jaws)
Context	S1 : Which was the best scene in the movie according to you? S2 : I liked the one in which Rocky trains and runs. S1 : It was great when the match commences, and although Donnie takes a few hits, he defeats Leo. S2 : Creed goes the distance as an exceptional crowd pleaser. S1 : I agree with grit, style and substance. There's no other way to say it; Creed is a knockout from start to finish.	S2 : The scene where they're yelling at each other out the window was gold S1 : Thats a similar scene from family guy. I couldnt stop laughing. S2 : I thought every scene was gold. I also couldn't believe Liam Neeson was in the movie. S1 : I will love to pick something similar this evening. Do you have a recommendation?	S1 : What do you think about the characters in this movie? S2 : My favorite character was Brody, he was such a fun character. S1 : I didn't get Quint's backstory.
Ground Truth	It wholeheartedly captures what was so brilliant about the first film: the characters.	Yes. You can try 22 Jump Street.	Quint, aboard that ship in World War II when it was sunk by torpedoes, had witnessed the deaths of roughly 800 men that day, many of whom were eaten by sharks as they struggled helplessly in the water.
HRED	yeah , but as the ark is launched , and you are right taxi driver , runs with the beauty , <UNK><UNK>and <UNK><UNK>.	the plot is a classic in the film , it 's not that it 's an original and cinematic experience .	i 'm not ashamed to admit it , mary poppens is bangin ! ! !
BiDAF(o)	I can't say enough great things about writer / director Ryan Coogler.	The Hangover	Hooper notices the shark beginning to circle the boat, and Quint rushes out for a look.
BiDAF(ms)	I loved how they introduced each opponent with his record and accomplishments.	21 Jump Street	The mood suddenly darkens as Quint admits that the former tattoo represented the US Navy cruiser
GTTP(o)	it wholeheartedly captures what was so brilliant about the first film : the characters .	$ 218,628,680 i think .	brody is stunned and alerts quint .
GTTP(ml)	i can't say enough great things about writer / director ryan coogler .	this movie had way too much product placement .	jaws' is the original summer blockbuster

Table 2: Examples produced by various models. S1 denotes Speaker1 while S2 denotes Speaker2. 'o','ml' and 'ms' represent *oracle*, *mixed-long* and *mixed-short* versions of the dataset respectively.



Instructions

Imagine you are chatting with a friend about the movie ${movie_name}. For the given conversation , **fill in appropriate response by picking up a sentence from the given passage so that it fits in with the conversation**. Wikipedia page about the movie can be found here
Pleae keep the following rules in mind

1. You need to provide **three appropriate responses.**
2. If there are two passage, try to pick one response each from either of them.
3. You can pick full sentence or a part of the sentence but do not change the order of words in the given sentence.
4. You are free to add words at the start of the sentence or at the end of the sentence to improve the quality of the response.

Example : Interstellar

Document

Honestly, it was so beautiful that I felt like I was sucked into the movie. We can feel the talent of Christopher Nolan, just by looking at the way it is filmed. The techniques he used contribute to create that visual environment in a believable way. The sound environment is just mesmerizing.It is a very important part of the movie, because some scenes take place in space, and Nolan just found the right way to use sound. The soundtrack (made by the great Hans Zimmer) is breathtaking, epic, amazing, unreal. I could find a lot more adjectives to qualify it, but you have to hear it to understand how epic they are. These two important parts (image and sound) create a stunning atmosphere. You will forget you are in a movie theater, and you will be lost in space, sucked into the adventures of this new Space Odyssey, begging for more. It is a truly unique experience.

Chat

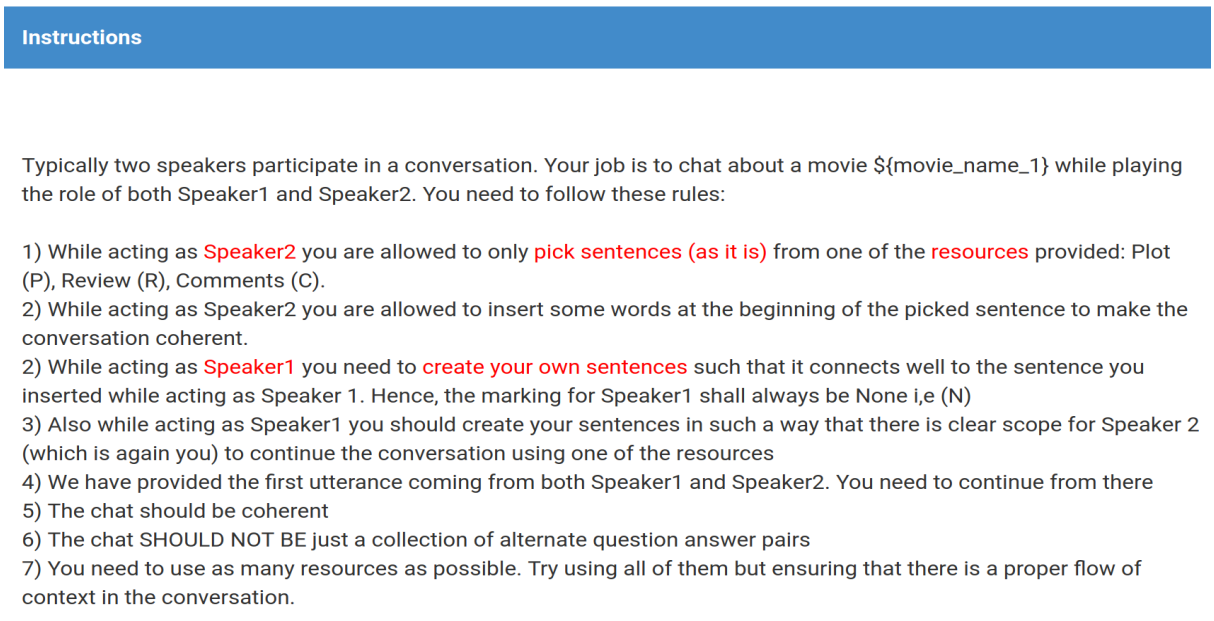
Speaker1 : What do you think about this movie ? Speaker2 : It is a well-crafted movie. Loved it! Speaker1 : Christopher Nolan is a genius film-maker

Responses

Response 1: We can feel the talent of Christopher Nolan, just by looking at the way it is filmed. Response 2: Such is the beauty of Nolan's direction. You will forget you are in a movie theater, and you will be lost in space. Response 3: Indeed.Nolan just found the right way to use sound. The background score was beautiful.

Notice how in Response1, Speaker2 continues the appreciating Chirstopher Nolan, in Response2 Speaker2 is a bit exaggerating about his experience and in Response3, Speaker2 apprecaites Nolan by highlighting the beauty of background score in the movie. All the above responses are apt in this conversation. The words in green are added to improve the quality of response

Figure 1: Instruction screen for collection of multiple-responses for the same chat for the test dataset.



Instructions

Typically two speakers participate in a conversation. Your job is to chat about a movie ${movie_name_1} while playing the role of both Speaker1 and Speaker2. You need to follow these rules:

1) While acting as Speaker2 you are allowed to only pick sentences (as it is) from one of the resources provided: Plot (P), Review (R), Comments (C).
2) While acting as Speaker2 you are allowed to insert some words at the beginning of the picked sentence to make the conversation coherent.
2) While acting as Speaker1 you need to create your own sentences such that it connects well to the sentence you inserted while acting as Speaker 1. Hence, the marking for Speaker1 shall always be None i,e (N)
3) Also while acting as Speaker1 you should create your sentences in such a way that there is clear scope for Speaker 2 (which is again you) to continue the conversation using one of the resources
4) We have provided the first utterance coming from both Speaker1 and Speaker2. You need to continue from there
5) The chat should be coherent
6) The chat SHOULD NOT BE just a collection of alternate question answer pairs
7) You need to use as many resources as possible. Try using all of them but ensuring that there is a proper flow of context in the conversation.

Figure 2: Instruction screen for chat data collection from Phase 4 of the dataset collection procedure

For example, consider the conversation on **Dunkirk**

Plot :

The British navy requisitions civilian vessels that can get close to the beach. In Weymouth, Mr. Dawson and his son Peter set out on his boat Moonstone rather than let the navy take it. Impulsively, their teenage friend George joins them. At sea, they rescue a shell-shocked officer from a wrecked ship. When he realises that Dawson is sailing for Dunkirk, the officer demands that they turn back, and tries to wrest control of the boat; in the struggle, George suffers a head injury that renders him blind.

Review :

Dunkirk is edge of your seat filmmaking. Can honestly say I've never seen anything like it.A lot of people were wondering about Harry_styles & unknown cast. They're all great but Dunkirk is not about any one solider. Also 'Dunkirk' is another brilliant collaboration between Nolan & HansZimmer. The way he mixes in a ticking clock with score is nail biting.DUNKIRK relies on very little dialogue.We all know what happened on that beach, but Nolan's take is worth visiting. Yes, DUNKIRK relies heavily on sound of an increasingly fast ticking clock to build suspense.Drop everything and go watch Dunkirk. It is an experience. Not a mere film.

Comment :

This is a very important movie, because it doesn't glamorize or glorify war.

I think the movie was brilliant .
Just awesome! Simply awesome!

Hans Zimmer did really great with the score and it really was an immersive experience

Fact Table :

Tagline	Hope is a weapon. Survival is victory. The event that shaped our world.
Similar Movies	Saving Private Ryan Interstellar The Sea Wolves

Figure 3: Background resources for the example chat shown to the workers on AMT

Conversation

Speaker1 (N) : What do you think about the movie ?
Speaker2 (C) : I think the movie was brilliant.
Speaker1 (N) : Agreed! One of the finest in this genre. (A casual sentence to start expressing your thoughts about the same.)
Speaker2 (C) : *I believe* the best part about the movie is that it doesn't glamorize or glorify war. (picked from a review - the words I believe are inserted at the beginning to make the conversation coherent)
Speaker1 (N) : Totally! Oh by the way do you remember the name of the ship headed by Mr.Dawson ? (Speaker1 listens and appreciates Speaker2's thoughts and continues discussing with a new aspect, in this case small details about the movie. Speaker 1 deliberately creates a question which can be answered from the plot)
Speaker2 (P) : Yes. It was Moonstone (This can be inferred only if you read the plot)
Speaker1 (N) : Right. I am always impressed by the Nolan - Hans Zimmer collaboration. (Speaker1 gets the answer, but there's no need of elaborating on Moonstone further. Hence Speaker1 talks about the people behind the movie - once again paving way for the next utterance to be picked up from the given resources)
Speaker2 (R) : The way he mixes in a ticking clock with score is nail biting. (Speaker2 continues appreciating Hans Zimmer.)
Speaker1 (N) : Some of the scenes seem so real, I feel I was present right there at that very moment. (Speaker1 creates a sentence which connects well to the previous utterance)
Speaker2 (R) : To sum it up, it is an experience. Not a mere film (Speaker 2 intelligently picks a sentence from the resources and gives a nice conclusion about his opinion on the film.)
Speaker1 (N) : That's an interesting way to put it. Can you recommend any other movies ?
Speaker2 (F) : You should check out Saving Private Ryan (Based on the "Similar Movies" from the fact table, Speaker 2 recommends another movie.)

Figure 4: Example chat shown to the workers on AMT

Instructions

Below you have been asked 4 questions about a specific movie. Your answer to each question should begin with a specific phrase as mentioned below.
You can refer to the wikipedia page of the movie here .

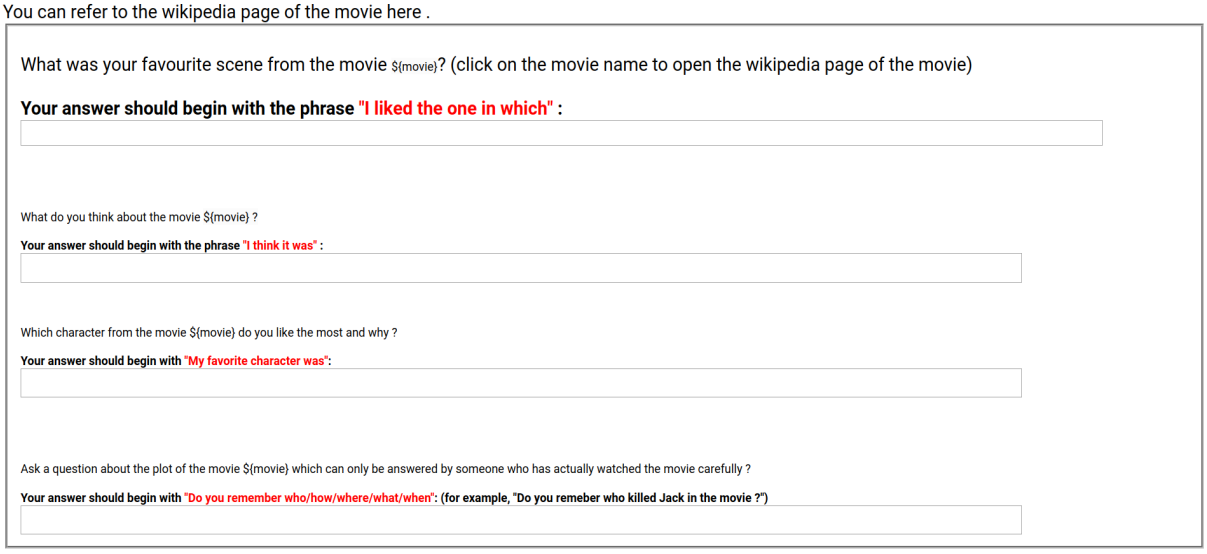


Figure 5: Instruction screen for opening statement collection from the stage of the dataset collection procedure