Modeling Multi-turn Conversation with Deep Utterance Aggregation

Zhuosheng Zhang*, Jiangtong Li*, Pengfei Zhu, Hai †, Gongshen Liu

Task Definition

- Each conversation in the concerned multi-turn response retrieval task can be described as a triple <C,R,Y>.
- $C = \{U_1, ..., U_t\}$ is the conversation context where $\{U_k\}$ denotes the k-th utterance.
- R is a response of the conversation.
- Y belongs to {0,1}, where $Y_i = 1$ means the response is proper, otherwise Yi = 0.
- The aim is to build a discriminator F (·, ·) on < C, R, Y >
- For each context-response pair {C, R}, F (C, R) measures the matching score of the pair.

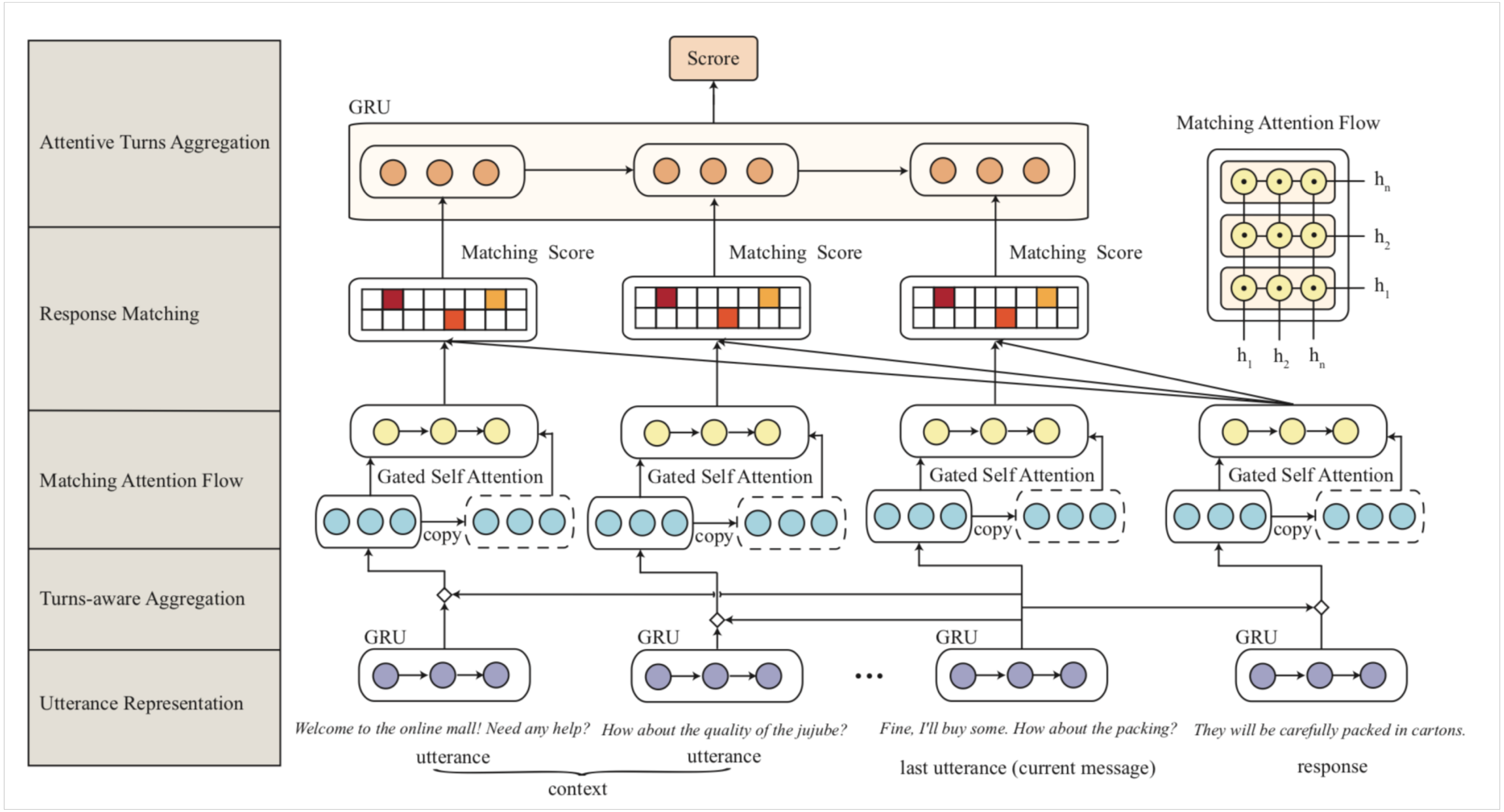
Motivation

- The relevance of each utterance to the supposed response usually varies.
- The last utterance in a conversation empirically conveys the user intention while the other utterances depict the conversation in different aspects.
- Words in an utterance also hold different importance to the whole utterance representation.

Contribution

- Use turns-aware aggregation to mix the last utterance with the previous ones.
- Employ self-attention based recurrent networks on each aggregated utterance.
- Release an E-commerce Dialogue Corpus (ECD) to facilitate the related studies.

Method


Attentive Turns Aggregation
Response Matching
Matching Attention Flow
Turns-aware Aggregation
Utterance Representation
Scrore
GRU
Matching Score
Matching Score
Matching Score
Matching Attention Flow
h_n
h_2
h_1
h_1 h_2 h_n
Gated Self Attention
Gated Self Attention
Gated Self Attention
Gated Self Attention
copy
copy
copy
copy
GRU
GRU
GRU
GRU
Welcome to the online mall! Need any help?
How about the quality of the jujube?
Fine, I'll buy some. How about the packing?
They will be carefully packed in cartons.
utterance
utterance
context
last utterance (current message)
response

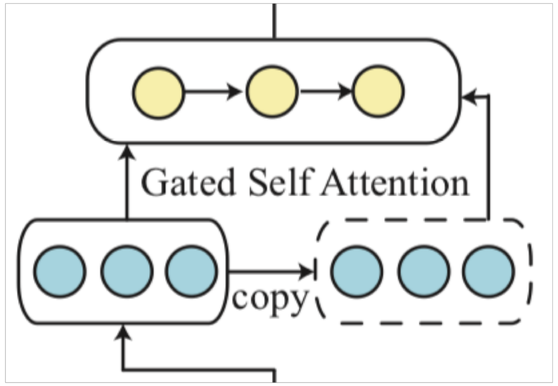
Method

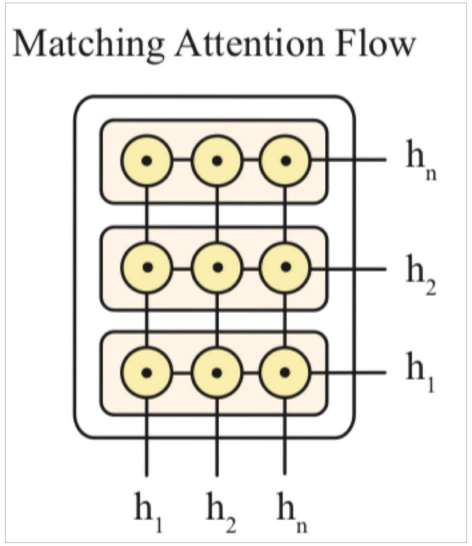
- Utterance Representation
  - Use GRU to encode each utterance and response respectively

$$z_i = \sigma(W_z u_i + V_z h_{i-1})$$
$$r_i = \sigma(W_r u_i + V_r h_{i-1})$$
$$\tilde{h_i} = tanh(W_h u_i + V_h(r_i \odot h_{i-1}))$$
$$h_i = z_i \odot \tilde{h_i} + (1 - z_i) \odot h_{i-1}$$

Method

- Turns-aware Aggregation
  - Mix the last utterance with the previous utterance and the response
- Matching Attention Flow
  - Using self-attention mechanism to filter the redundant information during the turns-aware aggregation





Method

- Response Matching
  - Calculate the matching matrix between every utterance and the response.
  - Use CNN to capture the correlation information for each utterance.

- Attentive Turns Aggregation
  - Use GRU to aggregate the correlation information in each utterance.

Dataset

- Ubuntu Dialogue Corpus
  - P:N = 1:1 for train
  - P:N = 1:9 for valid and test
- Douban Conversation Corpus
  - P:N = 1:1 for train and valid
  - P:N = 1:9 for test
  - More than one proper answer for test
- E-commerce Dialogue Corpus
  - Same as Ubuntu Dialogue Corpus

Results

Model	Ubuntu Dialogue Corpus			Douban Conversation Corpus					
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	P@1	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
TF-IDF	0.410	0.545	0.708	0.331	0.359	0.180	0.096	0.172	0.405
RNN	0.403	0.547	0.819	0.390	0.422	0.208	0.118	0.223	0.589
CNN	0.549	0.684	0.896	0.417	0.440	0.226	0.121	0.252	0.647
LSTM	0.638	0.784	0.949	0.485	0.537	0.320	0.187	0.343	0.720
BiLSTM	0.630	0.780	0.944	0.479	0.514	0.313	0.184	0.330	0.716
Multi-View	0.662	0.801	0.951	0.505	0.543	0.342	0.202	0.350	0.729
DL2R	0.626	0.783	0.944	0.488	0.527	0.330	0.193	0.342	0.705
MV-LSTM	0.653	0.804	0.946	0.498	0.538	0.348	0.202	0.351	0.710
Match-LSTM	0.653	0.799	0.944	0.500	0.537	0.345	0.202	0.348	0.720
Attentive-LSTM	0.633	0.789	0.943	0.495	0.523	0.331	0.192	0.328	0.718
Multi-Channel	0.656	0.809	0.942	0.506	0.543	0.349	0.203	0.351	0.709
Multi-Channel$_{exp}$	0.368	0.497	0.745	0.476	0.515	0.317	0.179	0.335	0.691
SMN	0.726	0.847	0.961	0.529	0.569	0.397	0.233	0.396	0.724
DUA	**0.752**	**0.868**	**0.962**	**0.551**	**0.599**	**0.421**	**0.243**	**0.421**	**0.780**

Results

Model	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
TF-IDF	0.159	0.256	0.477
RNN	0.325	0.463	0.775
CNN	0.328	0.515	0.792
LSTM	0.365	0.536	0.828
BiLSTM	0.355	0.525	0.825
Multi-View	0.421	0.601	0.861
DL2R	0.399	0.571	0.842
MV-LSTM	0.412	0.591	0.857
Match-LSTM	0.410	0.590	0.858
Attentive-LSTM	0.401	0.581	0.849
Multi-Channel	0.422	0.609	0.871
Multi-Channel$_{exp}$	0.352	0.556	0.827
SMN	0.453	0.654	0.886
DUA	**0.501**	**0.700**	**0.921**

Table 3: Comparison of different models on E-commerce Dialogue Corpus.

Ablation Study

	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
DUA	0.501	0.700	0.921
-CF	0.453	0.642	0.890
-MAF	0.432	0.625	0.883
-CF -MAF	0.413	0.613	0.867

Table 5: Ablation study on ECD dataset. CF and MAF denote the *Context Fusion* and *Matching Attention Flow*. The bracket means the context fusion approach adopted by the model.

Conclusion

- Propose a deep utterance aggregation approach to form a fine-grained context representation.
- Release the first e-commerce dialogue corpus to research communities.
- Experiments on three datasets show the model can yield new state-of-the-art results.

Thanks & QA