Sequential Matching Model for End-to-End Multi-Turn Response Selection

Qian Chen, Wen Wang
Speech Lab, DAMO Academy, Alibaba Group. {tanqing.cq, w.wang}@alibaba-inc.com

Alibaba Group 阿里巴巴集团

Contributions

★ Demonstrate that the potentials of **squenctial matching approaches** have not yet been fully exploited in the past for multi-turn response selection.

⋆ Previous state-of-the-art models used hierarchy-based (utterance-level and token-level) neural networks to explicitly model the interactions among the different turns' utterances for context modeling

★ The proposed models achieve new **state-of-the-art** performances on two large-scale public multi-turn response selection benchmark datasets.

Hierarchy-based vs Sequence-based

★ Hierarchy-based

+ model the multi-turn utterances' relationship **explicitly**
− **truncate** each utterances in the multi-turn context
  − if a large maximum length, increase computation complexity and memory cost
  − if a small maximum length, throw away some important information

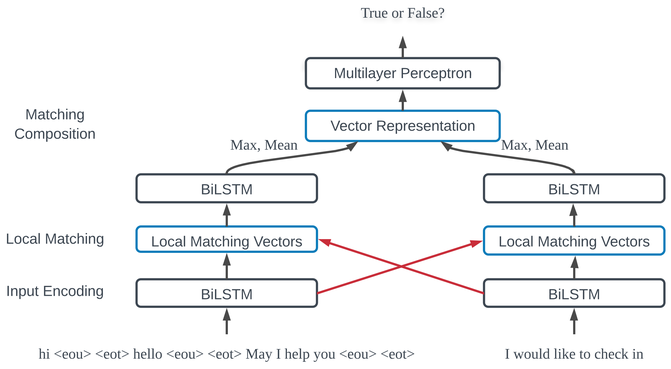
★ Sequence-based

− model the multi-turn utterances' relationship **implicitly**
+ **concatenate** the multi-turn context as a long sequence
  + it does not require each utterance to have the same length
  + lower computational complexity and memory cost

An example of multi-turn response selection

User	Utterance in Context
U1	hey guys, does your livecd have chroot installed? and bash?
U2	sure
U1	does it have everything I need to format a partition ext2?. and ext3?
U2	yep
U1	yay I can use it to install gentoo. !
U2	lol. LOL
U1	=-). brb rebooting into ubuntu
U2	form last week:. 04:21:47] <findme> this is a big crowd here. [04:21:53] <findme> have all gentoo users moved here ?
U1	to bad its still using apt I would switch in a heart beat if it had its own package manager
Candidate Responses	
1.	issues with msn?. I'm experiencing them on windows atm, current msn version
2.	what are you missing in apt ?
...	
10.	lspci will list your hardware, take a look at the VGA line
Answer	
	what are you missing in apt ?

Proposed model - based on ESIM



Model description in detail

1. Input Encoding

Context: $\boldsymbol{a} = (a_1, \ldots, a_m)$
Candidate Response: $\boldsymbol{b} = (b_1, \ldots, b_n)$

$$\boldsymbol{a}_i^s = \mathrm{BiLSTM}(\mathbf{E}(\boldsymbol{a}), i)\,, \quad (1)$$
$$\boldsymbol{b}_j^s = \mathrm{BiLSTM}(\mathbf{E}(\boldsymbol{b}), j)\,. \quad (2)$$

2. Local Matching

$$e_{ij} = (\boldsymbol{a}_i^s)^{\mathrm{T}} \boldsymbol{b}_j^s\,. \quad (3)$$

Token pairs with semantic relationship are probably aligned together.

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n}\exp(e_{ik})}\,,\ \boldsymbol{a}_i^c = \sum_{j=1}^{n}\alpha_{ij}\boldsymbol{b}_j^s\,, \quad (4)$$

$$\beta_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{m}\exp(e_{kj})}\,,\ \boldsymbol{b}_j^c = \sum_{i=1}^{m}\beta_{ij}\boldsymbol{a}_i^s\,, \quad (5)$$

$$\boldsymbol{a}_i^m = G([\boldsymbol{a}_i^s; \boldsymbol{a}_i^c; \boldsymbol{a}_i^s - \boldsymbol{a}_i^c; \boldsymbol{a}_i^s \circ \boldsymbol{a}_i^c])\,, \quad (6)$$
$$\boldsymbol{b}_j^m = G([\boldsymbol{b}_j^s, \boldsymbol{b}_j^c; \boldsymbol{b}_j^s - \boldsymbol{b}_j^c; \boldsymbol{b}_j^s \circ \boldsymbol{b}_j^c])\,, \quad (7)$$

Through comparing $\boldsymbol{a}_i^s$ and $\boldsymbol{a}_i^c$, we can obtain token-level relationship information for each token.

3. Matching Composition

$$\boldsymbol{a}_i^v = \mathrm{BiLSTM}(\boldsymbol{a}^m, i)\,, \quad (8)$$
$$\boldsymbol{b}_j^v = \mathrm{BiLSTM}(\boldsymbol{b}^m, j)\,. \quad (9)$$

The final vector are fed to MLP classifier.

$$y = \mathrm{MLP}([\boldsymbol{a}_{max}^v; \boldsymbol{a}_{mean}^a, \boldsymbol{b}_{max}^v; \boldsymbol{b}_{mean}^v])\,. \quad (10)$$

Experiment set-up

- **Dataset:** Ubuntu (**English**) and E-commerice (**Chinese**)
- **Metrics:** Recall at position k (R@k), i.e., R@1, R@2 and R@5
- **Training Detail:** word2vec embedding, Adam, hidden size 300

Name	Ubuntu			E-commerce		
	Train	Dev	Test	Train	Dev	Test
# context-response pairs	1M	500K	500K	1M	10K	10K
# candidates per context	2	10	10	2	2	10
Ave.# tokens of context	135	134	135	49	49	51
Ave.# tokens of response	21	21	21	7	7	10
Vocabulary size	180K	180K	440K	36K	10K	6K

Results

★ Our proposed ESIM sequential matching model outperformed all previous results.

★ Last few utterances in context are more important than the first few utterances.

Table 1: Comparison of different models on two benchmark datasets.

Models	Ubuntu			E-commerce		
	R@1	R@2	R@5	R@1	R@2	R@5
TF-IDF	0.410	0.545	0.708	0.159	0.256	0.477
CNN	0.549	0.684	0.896	0.328	0.515	0.792
BiLSTM	0.630	0.780	0.944	0.355	0.525	0.825
MV-LSTM	0.653	0.804	0.946	0.412	0.591	0.857
Match-LSTM	0.653	0.799	0.944	0.410	0.590	0.858
Attentive-LSTM	0.633	0.789	0.943	0.401	0.581	0.849
Multi-Channel	0.656	0.809	0.942	0.422	0.609	0.871
Multi-View	0.662	0.801	0.951	0.421	0.601	0.861
DL2R	0.626	0.783	0.944	0.399	0.571	0.842
SMN	0.726	0.847	0.961	0.453	0.654	0.886
DUA	0.752	0.868	0.962	0.501	0.700	0.921
DAM	0.767	0.874	0.969	-	-	-
Proposed ESIM	**0.796**	**0.894**	**0.975**	**0.570**	**0.767**	**0.948**

Table 2: Ablation over ESIM model on Ubuntu dataset. **CtxLen** = maximum length of context; **RepLen** = maximum length of response; **Rev** = truncate the context in reverse direction. **Emb** = the type of pre-trained word embedding.

Hyperparams				Dev Result		
CtxLen	RepLen	Rev	Emb	R@1	R@2	R@5
400	150	Y	W2V	0.797	0.893	0.976
400	150	Y	**Fasttext**	0.776	0.876	0.970
400	150	Y	**Random**	0.732	0.844	0.958
300	150	Y	W2V	0.793	0.892	0.976
200	150	Y	W2V	0.793	0.891	0.976
100	150	Y	W2V	0.783	0.886	0.974
400	**100**	Y	W2V	0.795	0.893	0.976
400	**50**	Y	W2V	0.792	0.892	0.975
400	150	**N**	W2V	0.793	0.892	0.976
100	150	**N**	W2V	0.707	0.827	0.951