Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots

Yu Wu
SKLSDE, Beihang University
wuyu@buaa.edu.cn

Wei Wu
Microsoft Corporation
wuwei@microsoft.com

Zhoujun Li
SKLSDE, Beihang University
lizj@buaa.edu.cn

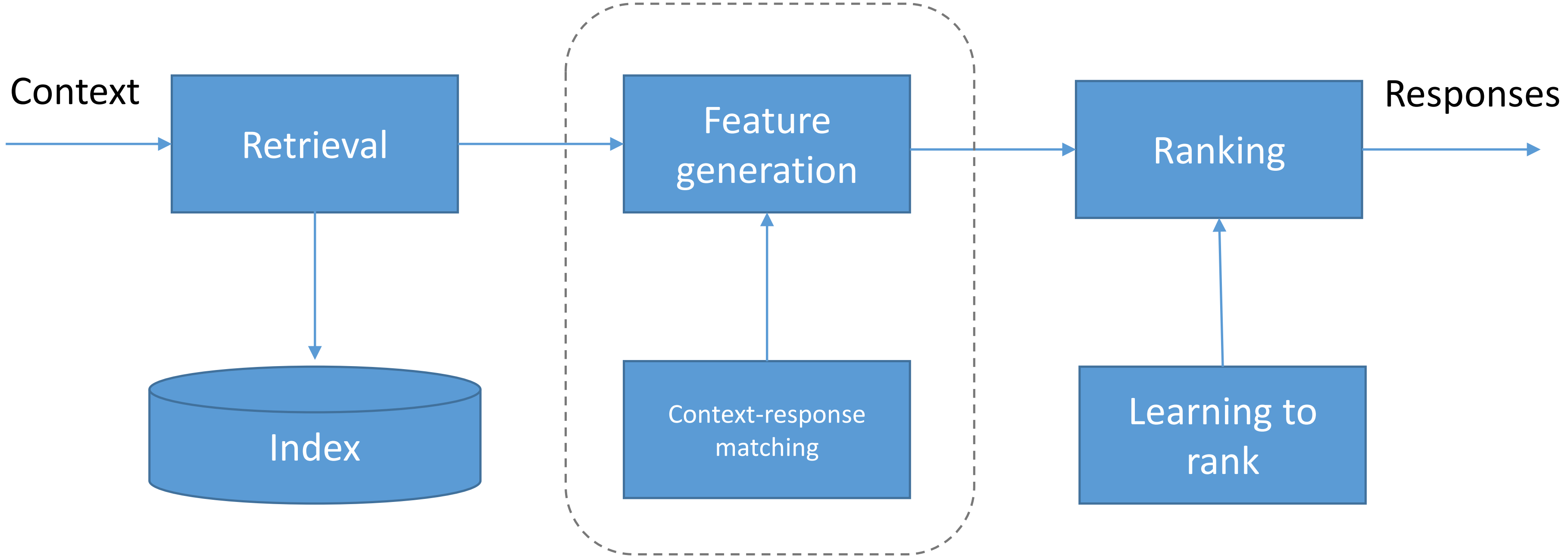
Ming Zhou
Microsoft Research
mingzhou@microsoft.com

Outline

- Task, challenges, and ideas
- Our approach
  - A new learning method for matching models.
- Experiment
  - Datasets
  - Evaluation and analysis

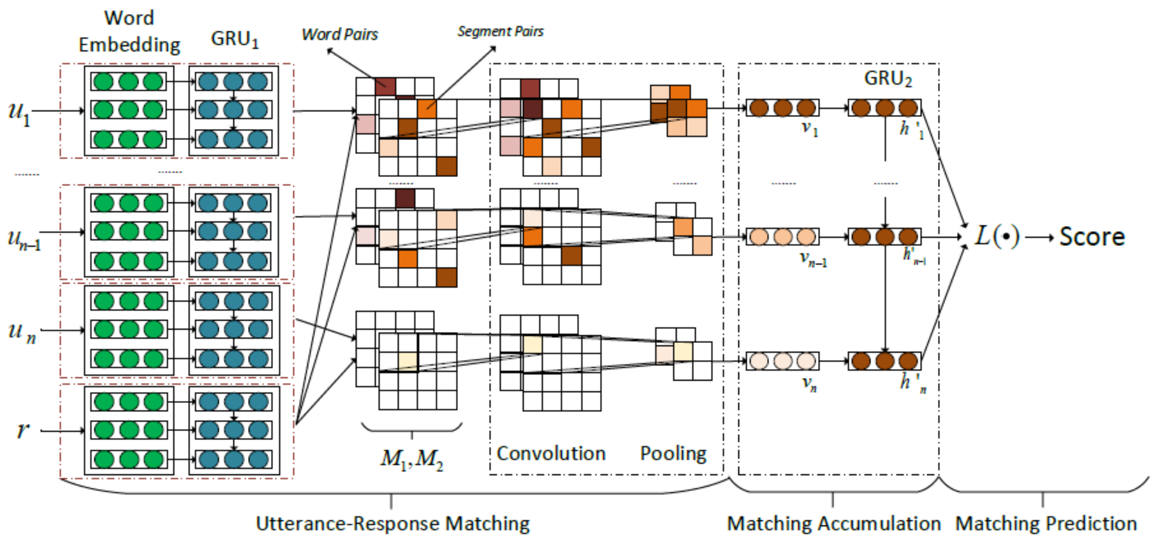
Task: retrieval-based chatbots

- Given a message, find most suitable responses
- Large repository of message-response pairs
- Take it as a search problem



Related Work

- Previous works focus on network architectures.
  - Single Turn
    - CNN, RNN, syntactic based neural networks ….
  - Multiple Turn
    - CNN, RNN, attention mechanism…
- These models are data hungry, so they are trained on large scale negative sampled dataset.

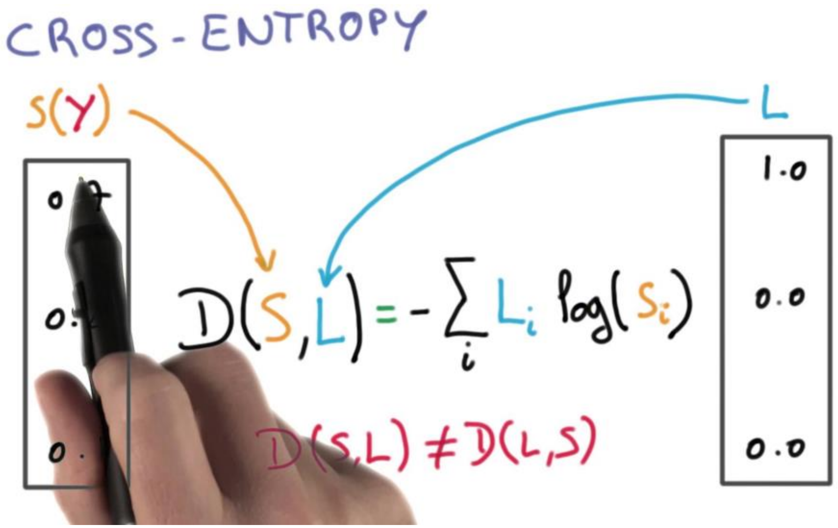


State-of-the-art multi-turn architecture (Wu et al. ACL 2017)

Background-----Loss Function

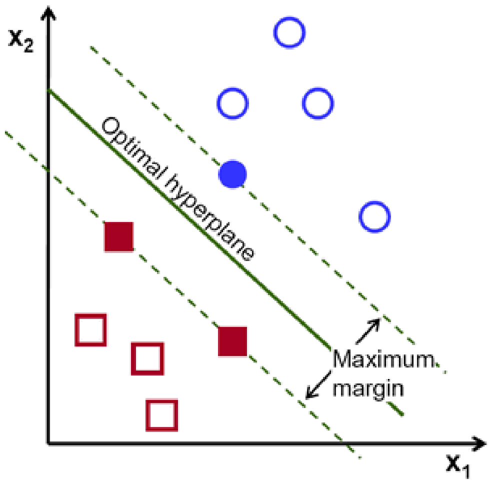
Cross Entropy Loss (Pointwise loss)

- $L = -\sum_i \mathrm{p}_i \log(\widehat{p_i})$



Hinge Loss (Pairwise loss)

- $S(+) - S(-) > \varepsilon$
- $L = \max(0, S(-) - S(+) + \varepsilon)$



Background: traditional training method

Given a (Q,R) pair, we first randomly sampled N instances $\{(Q, R_i^-)\}_N$.

Update the designed model with the use of point-wise cross entropy loss.

Test model on human annotation data.

Two problem:

1. Most of the randomly sampled responses are far from the semantics of the messages or the contexts.
2. Some of randomly sampled responses are false negatives which pollute the training data as noise.

Challenges of Response Selection in Chatbots

- Negative sampling oversimplifies response selection task in the training phrase.
  - Train: Given a utterance, positive responses are collected from human conversations, but negative ones are negative sampled.
  - Test: Given a utterance, a bunch of responses are returned by a search engine. Human annotators are asked to label these responses.

- Human labeling is expensive and exhausting, one cannot have large scale labeled data for model training.

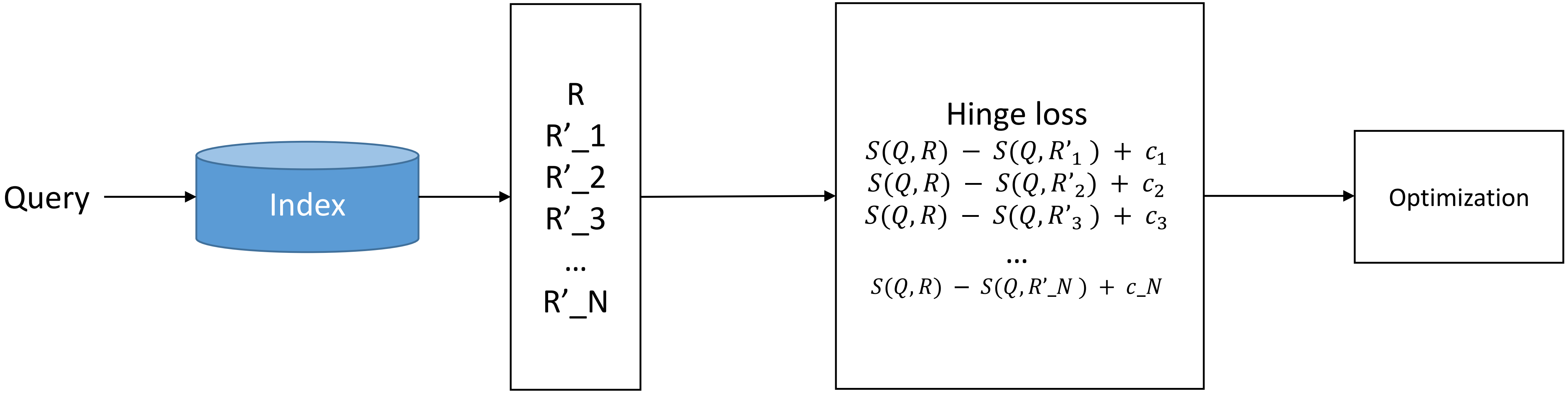
Outline

- Task, challenges, and ideas
- Our approach
  - A new learning method for matching models.
- Experiment
  - Datasets
  - Evaluation and analysis

Our Idea

Out training process

The margin in our loss
is dynamic.



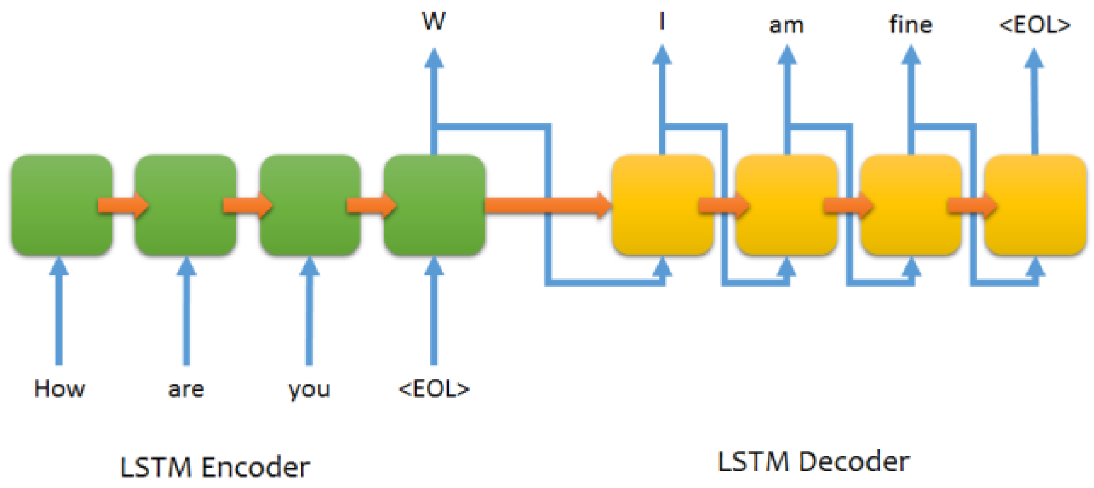
R is the ground-truth response,
and R’_i is a retrieved instance.

C_i is a confidence score for each
instance.
Our method encourages the model
to be more confident to classify a
response with a high c_i as a
negative one.

How to calculate the dynamic margin?

- We employ a Seq2Seq model to compute c_i.
  - Seq2Seq model is a unsupervised model.
  - It is able to compute a conditional probability likelihood $P(R|Q)$ without human annotation.

- $c_i = \max(0, \frac{s2s(Q,R_i)}{s2s(Q,R)} - 1)$



A new training method

Pre-train the matching model with negative sampling and cross entropy loss.

→ Given a (Q,R) pair, retrieve N instances $\{(Q, R_i^-)\}_N$ from a pre-defined index.

→ Update the designed model with the dynamic hinge loss.

→ Test model on human annotation da

The pre-training process enables the matching model to distinguish semantically far away responses.

1. Oversimplification problem of the negative sampling approach can be partially mitigated.
2. We can avoid false negative examples and true negative examples are treated equally during training

Outline

- Task, challenges, and ideas
- Our approach
  - A new learning method for matching models.
- Experiment
  - Datasets
  - Evaluation and analysis

Dataset

- STC data set (Wang et al., 2013)
  - Single-turn response selection
  - Over 4 million post-response pairs (true response) in Weibo for training.
  - The test set consists of 422 posts with each one associated with around 30 responses labeled by human annotators in "good" and "bad".
- Douban Conversation Corpus (Wu et al., 2017)
  - Multi-turn response selection
  - 0.5 million context-response (true response) pairs for training
  - In the test set, every context has 10 response candidates, and each of the response has a label "good" or "bad" judged by human annotators.

Evaluation Results

	P@1
TFIDF (Wang et al., 2013)	0.574
+Translation (Wang et al., 2013)	0.587
+WordEmbedding	0.579
+DeepMatch$_{topic}$ (Lu and Li, 2013)	0.587
+DeepMatch$_{tree}$ (Wang et al., 2015)	0.608
+LSTM (Lowe et al., 2015)	0.592
+LSTM+WS	0.616
+CNN (Hu et al., 2014)	0.585
+CNN+WS	0.604

Table 1: Results on STC

	MAP	MRR	P@1
TFIDF	0.331	0.359	0.180
RNN	0.390	0.422	0.208
CNN	0.417	0.440	0.226
BiLSTM	0.479	0.514	0.313
DL2R (Yan et al., 2016)	0.488	0.527	0.330
LSTM (Lowe et al., 2015)	0.485	0.527	0.320
LSTM+WS	0.519	0.559	0.359
Multi-View (Zhou et al., 2016)	0.505	0.543	0.342
Multi-View+WS	0.534	0.575	0.378
SMN (Wu et al., 2017)	0.526	0.571	0.393
SMN+WS	0.565	0.609	0.421

Table 2: Results on Douban Conversation Corpus

Ablation Test

- +WSrand: negative samples are randomly generated.
- +const: the marginal in the loss function is a static number.
- +WS: Our full model

	STC	Douban		
	P@1	MAP	MRR	P@1
CNN+WSrand	0.590	-	-	-
CNN+const	0.598	-	-	-
CNN+WS	0.604	-	-	-
LSTM+WSrand	0.598	0.501	0.532	0.323
LSTM+const	0.607	0.510	0.545	0.331
LSTM+WS	0.616	0.519	0.559	0.359
Multi-View+WSrand	-	0.515	0.549	0.357
Multi-View+const	-	0.528	0.564	0.370
Multi-View+WS	-	0.534	0.575	0.378
SMN+WSrand	-	0.536	0.574	0.377
SMN+const	-	0.558	0.603	0.417
SMN+WS	-	0.565	0.609	0.421

Table 3: Ablation results.

More Findings

- Updating the Seq2Seq model is not beneficial to the discriminator.
- The number of negative instances is an important hyper-parameter for our model.

	$LSTM_2$	$LSTM_5$	$LSTM_{10}$	$LSTM_{20}$
P@1	0.603	0.608	0.615	0.616
	SMN_2	SMN_5	SMN_{10}	SMN_{20}
MAP	0.542	0.556	0.565	0.567
MRR	0.588	0.594	0.609	0.609
P@1	0.408	0.412	0.421	0.423

Table 4: The effect of instance number

Conclusion

- We study a less explored problem in retrieval-based chatbots.
- We propose of a new method that can leverage unlabeled data to learn matching models for retrieval-based chatbots.
- We empirically verify the effectiveness of the method on public data sets.