splab at the NTCIR-13 STC-2 Task

Xuan Liu
Shanghai, China
1097114522@sjtu.edu.cn

Xueyang Wu
Shanghai, China
wuxueyang@sjtu.edu.cn

Ruinian Chen
Shanghai, China
ruinian.chen@sjtu.edu.cn

Zijian Zhao
Shanghai, China
1248uu@sjtu.edu.cn

Hongtao Lin
Shanghai, China
Linkin_HearT@sjtu.edu.cn

Kai Yu
Shanghai, China
kai.yu@sjtu.edu.cn

ABSTRACT

The splab team participated in the Chinese subtask of the NTCIR-13 on Short Text Conversation(STC-2)[9] Task. A large amount of pairs of post-comments are provided as data repository or training set. Given a new post, we need to retrieve suitable responses from the existing comments(retrieval-based method) or generate appropriate replies with the model trained on the data(generation-based method). We adopt the generation-based method and develop several systems based on the encoder-decoder framework. In our systems, we first implement the basic encoder-decoder architecture enhanced by attention mechanism. However this model tends to generate short and dull responses. And to alleviate this problem, we utilize two ways to enrich the information contained in the generated responses. One way is multiresolution recurrent neural network and the other is sequence generation based on word-level external memory. Both methods focus on some keywords in the post so as to generate targeted and informative responses. And we also use a DSSM to rerank the candidates to select the most reasonable responses. The evaluation of submitted results do not correspond with what we have expected and we believe that it reflects the weakness of human subjective evaluation.

Team Name

splab

Subtasks

Short Text Conversation (Chinese)

Keywords

short text conversation, sequence-to-sequence learning, multiresolution, word-level external memory, DSSM

1. INTRODUCTION

The Speech Lab of Shanghai Jiao Tong University participated in the NTCIR-13 Short Text Conversation Task [5]. This task aims to tackle the problem of one-round dialogue given a large repository of post-comment pairs. This task contains two subtasks, i.e. , Chinese subtask and Japanese subtask. It also recommends two methods, namely, retrieval-based method or generation-based method. Our work focus on the Chinese subtask with generation techniques.

Multi-round interactive dialogue has always been a heated topic and a milestone in the field of natural language processing. Short text Conversation can be seen as a simplified version of the task as it has only single-round. And recently it is promising to generate natural sentences using the deep learning techniques, especially the recurrent neural networks [12]. And the encoder-decoder framework has been applied to machine translation[2], conversational modeling[8] and so on.

In the subtask, we build several systems which sharing the same encoder-decoder framework. We first implement the basic system called neural responding machine[8]. Although the model can generate fluent sentences, these responses tend to be short and dull. And we notice that these good comments always correspond to a specific word in the post. So to enrich the information in the generated responses, we further more develop two word-based sequence generation models.

The multiresolution recurrent neural network first predicts keywords sequence and then takes it as additional input besides besides post to generate natural language responses.

And another method just select some words which are always adjectives as keywords. And the keyword is represented as a cluster of vectors of sentences containing the word. The vector is coupled with the encoded vector of post to generate replies.

Finally, we also develop a reranking mechanism which makes use of a DSSM to rerank all the generated responses.

The remainder of this paper is organized as follows. Section 2 describes our system in detail. Our submitted results and some discussion are presented in Section 3. We conclude in Section 4.

2. OUR SYSTEM

In this section, we will introduce the three models, namely , basic encoder-decoder architecture equipped with attention mechanism[6], multiresolution recurrent neural network[7] and sequence generation method with word-level external memory. And we will also present the reranking mechanism.

2.1 Data Preprocessing

In order to enlarge the training corpus, we mix the repository of stc1 and stc2 and delete the overlapping part. We segment the posts and responses by LTP [1]. After that, we delete post or comment with a length of more than 30 words to reduce the time span of the model. We also delete the sentence which contains more than one rare word. Since one post may correspond to several responses, in order to avoid over-emphasizing some posts, we also truncate the post- response pair if the corresponding post appears more than 100

times. We split the remaining data into training set and validation set. Training set and validation set share no posts. we construct the vocabulary independently for post and response. Any words not in the vocabulary are replaced by a special symbol UNK.

2.2 Encoder-Decoder Architecture for Conversation

The encoder-decoder framework is first proposed to tackle the problem of machine translation[2]. But it is quickly applied to other tasks include conversational modeling[8]. Our basic system barely implement the neural responding machine[8] with little difference in parameter settings and attention mechanism.

In sequence-to-sequence generation tasks, each input X is paired with a sequence of outputs to predict: $Y = y_1, y_2, ..., y_n$. Here, the last token y_n is EOS, which means the end of sentence. The networks defines a distribution over outputs and sequentially predicts tokens using a softmax function.

$$P(Y|X) = \prod_{t=1}^{t=n} p(y_t|x_1, x_2, ..., x_m, y_1, y_2, ..., y_{t-1})$$

The Goal of maximum likelihood training is minimizing:

$$-\sum_{i=1}^{n} \log P(Y_i|X_i) =$$

$$-\sum_{i=1}^{n}\sum_{t=1}^{l_i} \log p(y_t^i|x_1^i, x_2^i, ..., x_{m_i}^i, y_1^i, y_2^i, ..., y_{t-1}^i)$$

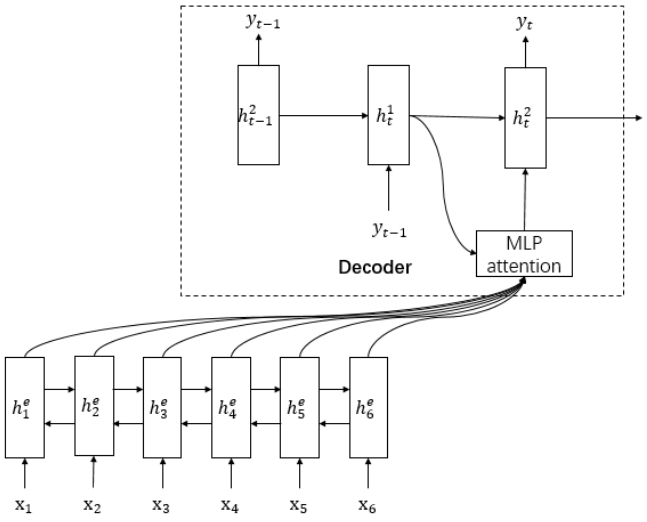

Figure 1: Network Architecture

The encoder is a one-layer bidirectional LSTM. We concatenate the last hidden layer of each direction as the initial hidden vector of the decoder. Traditional sequence-to-sequence model encodes the information of the post into a fixed dimension vector, which cannot encode sufficient information when the post is long. To solve this issue, attention mechanism is introduced. We also apply this method. We denote the hidden vector and cell vector of encoder as h_t^e, c_t^e. The decoder has two LSTM Cell. However, it connect in series, rather than in parallel. The hidden vector and cell vector of each LSTM is denoted as h_t^1, c_t^1, h_t^2, c_t^2. (The initial hidden vector of decoder is h_0^2).

First, we compute the hidden vector of first LSTM Cell.

$$h_t^1, c_t^1 = f^1(y_{t-1}, h_{t-1}^2, c_{t-1}^2)$$

After that, we add attention information. The context vector of time step t is:

$$ctx_t = \sum_{1}^{t} \alpha_t h_t^e$$

The attention score is computed by a two layer forward network, which is denoted as g. The input of g is the hidden vector of encoder in each time step h_t^e and the first hidden vector of decoder in this step. The output is a scalar representing the importance of each time step in encoder. Then, the score of each time step is fed to a softmax-layer to compute the alignment.

$$a_t = \frac{\exp^{g(h_t^1, h_t^e)}}{\sum_1^t \exp^{g(h_t^1, h_t^e)}}$$

After computing the context vector ctx_t, the second hidden vector of the decoder is:

$$h_t^2, c_t^2 = f^2(ctx_t, h_t^1, c_t^1)$$

Finally, we calculate the probability of generating next token.

$$p(y_t|x_1, x_2, ..., x_m, y_1, y_2, ..., y_{t-1}) = \frac{\exp(w_{y_t} h_t^2 + b_{y_t})}{\Sigma_y \exp(w_y h_t^2 + b_y)}$$

2.3 Multi-resolution Recurrent Neural Network

we also implement multi-resolution [7] method, which consists of two steps. The first step uses a sequence-to-sequence network to predict keywords sequence. The second step Taking keywords sequence as additional input besides post to another sequence-to-sequence network to generate the nature language response. During training, the keywords sequence input is the ground truth. During generation, the keywords sequence is generated according to the first step.

Different from [7], The POS of golden keywords sequence is not limited to Noun. Noun, Verb and Adjective can be keywords unless it is in stopword list. In order to make the keyword sequence longer, the length of the generated keywords sequence is at least two.

In this experiment, we also create a smaller dataset according to the mutual information between the keywords of post and comment. The purpose of doing this operation is to delete the training data that is not closely related. G5 is the result of the original dataset and G1 is the result of the smaller dataset.

2.4 Sequence Generation with Word-Level External Memory

We observe that good comments are always related to some keywords in the post. So we would like to establish the coupling between post and keywords to generate meaningful responses. However, a single word lacks in semantics and can not provide enough information. Inspired the thought that "a word can be represented by the words around it", we think that a sentence containing some word reflects the semantics of the word in a specific context. We represent a keyword as a cluster of sentence vectors where every sentence contains

the keyword. When generating responses for a post, we first select a keyword and index the corresponding vector. Then we utilize both encoded vector of the post and vector of the keyword to decode the comments

2.5 Reranking Mechanism

We use a CNN-based deep semantic similarity model(C-DSSM) for comment reranking. The DSSM has wide applications including information retrieval and web search ranking [10, 3, 11]. The architecture of C-DSSM is illustrated in figure 2. The C-DSSM contains encoder which encode sentences into semantic feature vectors, and similarity measurement parts which measure the similarity(a real number) between two semantic feature vectors. In our experiments, we use CNNs as the encoder, this kinds of CNNs proposed in [4] are wildly used in text classification. Despite the CNN with one layer of convolution is simple, this model achieves excellent results in many benchmarks. Here we simply use cosine distance for similarity measurement.

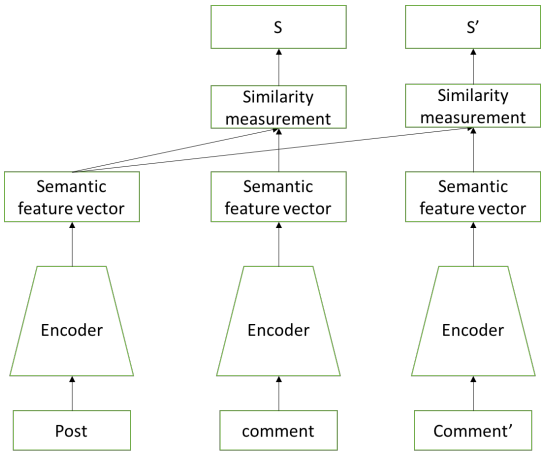

Figure 2: The architecture of C-DSSM

In the training phase, we feed the C-DSSM with three sentences which are post, ground truth comment and fake comment. The fake comment is randomly sampled from all comments. We obtain the semantic feature vectors of each sentence by the encoders. Note that, all of those three encoders are sharing weights. Then we calculate the similarity between post and ground truth comment $S(p, c)$, and similarity between post and fake comment $S(p, c')$. To train the model, we minimize a margin-based ranking criterion over the train set:

$$L = \sum_{p,c,c'} [\gamma + S(p, c) - S(p, c')]_+ \quad (1)$$

where $[x]_+$ denotes the positive part of x, and $\gamma > 0$ is a margin hyperparameter.

In the reranking phase, we feed the C-DSSM with post-comment pairs and use the similarity measurement scores for comment reranking. The lower score indicates higher relativity of the post-comment pair.

3. SUBMITTED RESULTS

In Chinese short text conversation task, we totally submit five runs. The evaluation results are listed in Table 1. splab-C-G1 and splab-C-G5 uses the multiresolution recurrent neural networks. splab-C-G2 utilize the sequence generation method with word-level external memory. splab-C-G4 is the basic encoder-decoder framework with attention mechanism. splab-C-G3 rerank all the comments generated by the previous four models and select the top-10 responses.

We observe in Table 1 that splab-C-G4 performs best among all systems. However, this result doesn't fit to our expectation. In our subjective evaluations which prefer diverse and informative responses , the other systems are always better than the basic system. We think it reveals that there is a lack of consistency between human subjective evaluations. And a objective criterion is required to alleviate this problem.

4. CONCLUSIONS

Interactive chatbots is an very interesting and challenging topic. Much effort has been devoted to the problem but there is little progress in this area. However, with the increasing amount of short text conversation data available online and development of deep learning techniques, we can try to tackle this problem.

In this task, we begin with the basic encoder-decoder framework. And to alleviate the problem of dull and short responses, we adopt word-based models to generate more informative responses, i.e. , multiresolution recurrent neural network and sequence generation method with word-level external memory. And the details has been fully introduced in the paper.

5. REFERENCES

[1] W. Che, Z. Li, and T. Liu. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics, 2010.

[2] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[3] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM, 2013.

[4] Y. Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

[5] S. Lifeng, S. Tetsuya, L. Hang, H. Ryuichiro, M. Yusuke, A. Yuki, and N. Masako. Overview of the ntcir-13 short text conversation task. In *Proceedings of NTCIR-13*, pages 1–100, 2017.

[6] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[7] I. V. Serban, T. Klinger, G. Tesauro, K. Talamadupula, B. Zhou, Y. Bengio, and A. C. Courville. Multiresolution recurrent neural networks: An application to dialogue response generation. In *AAAI*, pages 3288–3294, 2017.

Table 1: Evaluation results of Chinese short text conversation task for five runs

Run	Mean nG@1	Mean P+	Mean nERR@10
splab-C-G1	0.4903	0.5763	0.6416
splab-C-G2	0.4181	0.4877	0.5335
splab-C-G3	0.4481	0.5461	0.6084
splab-C-G4	0.5062	0.6016	0.6579
splab-C-G5	0.4536	0.5634	0.6278

[8] L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

[9] L. Shang, T. Sakai, H. Li, R. Higashinaka, Y. Miyao, Y. Arase, and M. Nomoto. Overview of the NTCIR-13 short text conversation task. In *Proceedings of NTCIR-13*, 2017.

[10] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. A latent semantic model with convolutional-pooling structure for information retrieval. CIKM, November 2014.

[11] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 373–374. ACM, 2014.

[12] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, 2011.