Contextual and Feature-based Models by PolyU Team at the NTCIR-13 STC-2 Task

[†]Yanran Li, [†,§]Hui Su,[†]Wenjie Li
[†]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[§]Institute of Software, Chinese Academy of Science, China
{csyli,cswjli}@comp.polyu.edu.hk suhui15@mails.ucas.ac.cn

ABSTRACT

The PolyU team participated in the Chinese Short Text Conversation (STC) subtask of the NTCIR-13, the core task of NTCIR-13. At NTCIR-13, generation-based approaches and their evaluations are firstly introduced into the task. This minority report describes our methods to solving the STC problem including four retrieval-based and two generation-based typical approaches. We compare and discuss the official results.

Keywords

retrieval-based conversational agents, rerank, generation-based conversational agents

Team Name

PolyU

Subtasks

Short Text Conversation Task (Chinese), a.k.a. STC-2 Task

1. INTRODUCTION

Conversational agents have emerged as new commercial channels and have received considerable attention from both industry and academia. They allow enterprises to reach customers in auxiliary platforms and interact with them in natural ways. To develop such intelligent agents involve series of critical natural language processing techniques including natural language representation, latent intention understanding, and natural language generation. These make research on machine conversations appealing yet challenging.

To evaluate approaches for conversational agents, researchers have collected and adopted a variety kinds of data. Some researchers gathered data from specific-domain and the dialogues they obtained are often task-oriented, which limited the potential of the trained agents [8, 7]. To acquire large and reliable open-domain conversation data is non-trivial due to the protection of user privacy. Instead, researchers seek alternatives from social platforms like Ubuntu and Reddit forums [5, 1]. On these platforms, users often involve in a thread focusing on a certain topic. The original posts are the following replies are often treated as conversation utterance pairs. Because these platforms contain millions of threads covering hundreds of topics, these kind of data are often used to train open-domain conversational models.

However, posts and replies on Ubuntu and Reddit forums are often too long to be captured in existing models. This hampers the research on developing powerful conversational models. To combat this issue, researchers often investigate data from Twitter and Chinese Weibo platform [9, 12]. These two platforms constrain each post and reply to be less than 140 words. This results in more compact meaning to be expressed in the posts and replies. In [12], Chinese Sina Weibo data has been collected for short text conversation task. The task is aimed to compare conversational models by human evaluation from different perspectives like relatedness and fluency.

The PolyU team participated in the Chinese Short Text Conversation (STC) subtask, the core task of NTCIR-13. In the past years, the STC subtask is only designed for retrieval approaches and the participated models are evaluated based solely on the relevance Information Retrieval (IR) measures. At NTCIR-13 this year, generation-based approaches has been encouraged and are evaluated independently from retrieval-based approaches. Teams are required to submit their results from these two distinguished kinds of approaches separately and the evaluation metrics are also different. An overview of the tasks and results in this year is provided by [10].

This minority report describes our methods to solving the STC-2 subtask including four retrieval-based (Section 4) and two generation-based typical approaches (Section 5). We compare and discuss the official results in Section 6.

2. CORPUS STATISTICS

We give a brief view of the corpus used in the subtask. The corpus is crawled from Chinese Sina Weibo platform and is officially separated into training and testing data. Note that data for the retrieval-based subtask and the generation-based subtask are given in two repositories.

In the retrieval-based repository, there are 219,174 Weibo posts and 4,305,706 corresponding replies, consisting of in total 4,433,949 post-replies pairs. Due to Weibo's mechanism, each post can have multiple replies (maybe from multiple users). Meanwhile, each reply is potentially appropriate to multiple posts. In the retrieval-based corpus, each post has in average 20 replies. The summary statistics of the corpus is given in Table 1.

The task also provides a portion of 769 labeled data where each post is associated with approximately 15 candidate replies. For each post as one query, replies are labeled as "suitable", "neural", and "unsuitable", resulting in 11,535 labeled post-reply pairs. Also, 100 posts are carefully selected as test data (test query) to evaluate submitted approaches. The summary of this part of data is in Table 2.

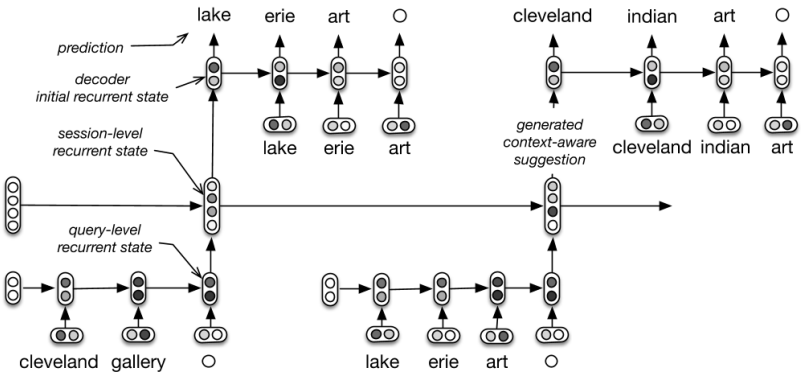

Figure 1: The overview of the hierarchical recurrent encoder-decoder (HRED) architecture proposed by [11]. The current utterance is encoded by a word-level EncoderRNN, and the dialogue history is stored in a utterance-level ContextRNN. The figure is directly adopted from [11].

Posts	219,174
Replies	4,305,706
Post-Reply Pairs	4,433,949
Average Replies Per Post	20

Table 1: Basic Statistics of Weibo Corpus in NTCIR-13 STC Task.

Labeled Post	769
Labeled Replies	11,535
Labeled Post-Reply Pairs	11,535
Test Post	100

Table 2: Summary Statistics of Labeled and Test data.

However, this portion of labeled data is artificially composed together. The replies are originally corresponding to the posts other than the query post. Additionally, the size of these labeled data is too small to train a neural conversational models. Thereafter, we omit this portion of labeled data during training and use them only to better understand the evaluation criteria, which will be briefly described in the following section.

3. EVALUATION MEASURES

The evaluation in STC task is conducted by three human assessors. They are required to rate each candidate post-reply pair by four criteria for both retrieval-based and generation-based approaches. The four criteria are designed to reflect the naturalness of the natural language in short text conversation. They are as follows:

- **Fluent** This measurement is designed to reflect whether the reply looks like natural language. To our understanding, this measurement is especially selected for generation-based approaches.

- **Coherent** This evaluates whether the reply is logically connected or typically related to the query post. In other words, this measure is designed to evaluate the semantic coherence between the candidate reply and the query post.

- **Self-sufficient** This measurement is newly introduced by NTCIR-13 and is aimed to encourage the candidate reply to be cohesively connected with the original post. The task interprets it as "the assessor can judge that the comment is appropriate by reading nothing other than the post-reply pair".

- **Substantial** This measurement indicates whether the candidate reply provides new information. To our understanding, this encourages the reply to be more informative and penalizes those generic ones, i.e., "Amazing!" and "I don't know".

Following these four criteria, if a candidate reply is fluent, coherent, self-sufficient, and substantial considering the query post, it will be labeled as 2. If it is fluent and coherent, but not self-sufficient and/or not substantial, it will be labeled as 1. Otherwise, it will be labeled as 0.

4. RETRIEVAL-BASED APPROACHES

In this section, we describe four retrieval-based approaches that we attempted in the task. All the approaches (including generation-based approaches) are based on hierarchical recurrent encoder-decoder (HRED) architecture [11] as shown in Figure 1.

To model the conversation history and the current utterance, HRED architecture comprises of three recurrent neural networks (RNNs). The current utterance is represented by a word-level EncoderRNN, and the conversation history is encoded in a utterance-level ContextRNN:

$$\mathbf{u}_t = \text{EncoderRNN}(t_1, \cdots, t_{n-1}) \quad (1)$$

$$\mathbf{c}_t = \text{ContextRNN}(w_1, \cdots, w_{t-1}) \quad (2)$$

where t denotes tokens in the utterance, and w denote history utterances. Based on the representations obtained from EncoderRNN and ContextRNN, the reply is decoded word-by-word by a DecoderRNN:

$$w_n \sim p_\theta(w_n | w_1, \cdots w_{n-1}) \quad (3)$$

For details, please refer to [11].

4.1 Basic Model

The first two retrieval models we tried are based on embedding representations. We measure the distance between embeddings as the average of cosine similarity, Jaccard distance and Euclidean distance.

In the first basic model, we compute the test post embeddings with those post embeddings in the training data. At test time, candidate replies whose post embeddings is closer to the test post embeddings are ranked higher.

4.2 Contextual Model

The second retrieval model we attempted is also based on embedding representations. Differently, we compute the test post embeddings not only with the post embeddings, but also with the corresponding reply embeddings. In other words, we take into consideration the semantic relatedness from both post and reply perspectives. This is inspired that the information in the original posts is often inadequate, and the replies sometimes provide supplementary and novel angel towards the same topic.

To collate the embeddings of posts and of replies, we have attempted two simple ways. We tried either average these two kinds of embeddings or concatenate these two embeddings, and found the former was the better. Therefore, we submitted the second retrieval-based approach with averaging techniques. At test time, candidate replies whose post and reply embeddings are closer to the test ones are ranked higher.

4.3 Feature-based Model

Other than word embeddings, we tried linguistic features in the third retrieval-based model. We adopted four main linguistic features: TF-IDF and three fuzzy string matching features, i.e., QRatio, WRatio, and Partial ratio. These fuzzy features is implemented with fuzzywuzzy package[1].

These features are used stationary. At the first state, we used TF-IDF to select 1,000 candidates for initial filtering. At the second state, we ranked these filtered 1,000 candidates with the three fuzzy features. These features have been proven effective on duplicate question detection task[2].

4.4 Reranking-based Model

The last approach we examined is reranking-based model. By introducing additional reranking stage, we are allowed to encourage the retrieved response to follow certain rules. Inspired by [6], we adopted a simple way to implement reranking models.

After investigating the labeled data provided by the task, we observed that typical emotional words are very beneficial factors in response selection. Hence, we manually built up a typical emotional lexicon and added them as special features in the reranking stage. We compare the emotional features of the test example with those of the candidate example, and used the compared similarity as reranking feature. For example, if the test query post is happy, then the candidate replies whose emotions are also positive will be reranked higher.

5. GENERATION-BASED APPROACHES

Our generation-based approaches are typically following the mainstream approaches. The simplest generation-based approach we adopted is a vanilla HRED with GRU as basic cell, as described before. Such approach is widely selected as baseline models in dialog generation [12, 5, 1].

We then tried the HRED architecture with attention mechanism [2] which has demonstrated its effectiveness on various NLP tasks including machine translation, dialog response generation, and reading comprehension.

6. RESULTS AND DISCUSSIONS

6.1 Experimental Setups

We tuned the parameters using 10-fold cross validation and submitted our results on test set. The word embeddings we used in all six approaches are set to 300-dimensional. They are initialized with Word2Vec embeddings trained on the Google News Corpus[3]. For the HRED architecture we adopted, the RNNs are 1-layer GRU with 512 hidden neurons [3]. The batch size was set as 128 and the learning rate was fixed as 0.0002. Models are optimized using Adam [4].

6.2 Task Evaluation Results

The task official evaluation hires three evaluators to give each candidate post-reply pair a grade of {0,1,2}. Then, the evaluation will be conducted by *gain values* and *unanimity-aware gain*. Instead of using the sum of labels as is, this method takes into account the agreements and differences among the three evaluators. We report the results of our approaches under these two measures.

	MSnDCG@0001	P-plus	nERR@0010
R1	0.0858	0.1649	0.1776
R2	0.1077	**0.2253**	**0.2387**
R3	<u>0.119</u>	<u>0.2117</u>	<u>0.2164</u>
R4	0.0542	0.1214	0.1125
G1	**0.1342**	0.1583	0.1239
G2	0.06	0.0814	0.0556

Table 3: Evaluation Results by Gain Values.

	MSnDCG@0001	P-plus	nERR@0010
R1	0.0929	0.17	0.1921
R2	0.1147	**0.2303**	**0.2516**
R3	<u>0.1255</u>	<u>0.2153</u>	<u>0.2289</u>
R4	0.0607	0.1273	0.1266
G1	**0.1407**	0.1605	0.1318
G2	0.0675	0.0849	0.0633

Table 4: Evaluation Results by Unanimity-aware Gain.

From Table 3 and Table 4 we can see that R2 and R3 models perform significantly better than other two retrieval based models. According to the descriptions in Section 4, they are contextual model and feature-based model. These results indicate two things: (1) the semantic information in the reply side is also informative and beneficial in response retrieval; (2) linguistic features like fuzzy string matching

[1]https://github.com/seatgeek/fuzzywuzzy

[2]https://github.com/abhishekkrthakur/is_that_a_duplicate_quora_question

[3]ttps://code.google.com/archive/p/word2vec/

features are powerful in semantic representation. We can easily explore these two findings in advanced approaches.

Surprisingly, attention-based generation approach perform quite terrible in this task. This is very different from a majority of existing research. We conjecture it as a failure to simply introducing attention mechanism into HRED architecture. In HRED, history information has been aggregated and summarized by ContextRNN, which is dynamic and might be controversial with simple attention mechanism. We leave it as future work to investigate deeper.

7. CONCLUSIONS

In this report, we describe our six approaches evaluated by NTCIR-13 Short Text Conversation Task. We attempt four retrieval-based approaches and two generation-based approaches based on sorts of features. The evaluation results suggest that the information in reply and linguistic features are beneficial for retrieval-based approaches.

8. REFERENCES

[1] R. Al-Rfou', M. Pickett, J. Snaider, Y.-H. Sung, B. Strope, and R. Kurzweil. Conversational contextual cues: The case of personalization and history for response ranking. *CoRR*, abs/1606.00372, 2016.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[3] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[4] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2014.

[5] R. Lowe, N. Pow, I. Serban, and J. Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL Conference*, 2015.

[6] C. Luo and W. Li. A combination of similarity and rule-based method of polyu for ntcir-12 stc task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, 2016.

[7] V. Petukhova, M. Gropp, D. Klakow, A. Schmidt, G. Eigner, M. Topf, S. Srb, P. Motlicek, B. Potard, J. Dines, et al. The dbox corpus collection of spoken human-human and human-machine dialogues. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC" 14)*, number EPFL-CONF-201766. European Language Resources Association (ELRA), 2014.

[8] E. K. Ringger, J. F. Allen, B. W. Miller, and T. Sikorski. A robust system for natural spoken dialogue. 1996.

[9] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics, 2011.

[10] 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 The NTCIR-13 Conference*, 2017.

[11] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. G. Simonsen, and J.-Y. Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Cikm'15 Proceedings of the 24th Acm International on Conference on Information and Knowledge Management*. Association for Computing Machinery, 2015.

[12] H. Wang, Z. Lu, H. Li, and E. Chen. A dataset for research on short-text conversations. In *EMNLP*, pages 935–945, 2013.