Gbot at the NTCIR-13 STC-2 Task

Hainan Zhang, Tonglei Guo, Yanyan Lan, Jiafeng Guo, Jun Xu, Jianing Li, Xueqi Cheng
1. CAS Key Lab of Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
2. University of Chinese Academy of Sciences, Beijing, China
{zhanghainan, guotonglei, lijianing}@software.ict.ac.cn,
{lanyanyan, guojiafeng, junxu, cxq}@ict.ac.cn

BACKGROUND

Problem:

For a given post X, the goal of STC-2 task is to generate or retrieve a response Y which looks like a real human response. The STC-2 task has two subtask: **generation-based method** and **retrieval-based method**. For generation-based subtask, we maximize the likelihood probability P(Y|X). For retrieval-based subtask, we return the top-k relevant comments with the given query.

Motivation:

For **generation-based method**, most existing works follow the Seq2Seq which is easy to generate general and safe responses [1,2]. We propose to add three kinds of constraint functions to the original Seq2Seq loss function, in order to guide the generation process.

For **retrieval-based method**, we apply method based on MatchPyramid, which first builds interactions and then uses a deep model to obtain the representation for the interactions and the relevance score.

GENERATION MODELS

Original Seq2Seq Loss Function

$$\mathcal{L} = -\sum_{(X,Y)\in\mathcal{D}} \log P(Y|X)$$

Our Constrained Loss Function

Constraint describing the quality of a generation G

$$\mathcal{L}_m = -\sum_{(X,Y)\in(D)} \boxed{\mathrm{cons_m}(X,G)} \times \log P(Y|X)$$

Three kinds of constraints

- Cosine Similarity function(SIM):

$$\mathrm{cons_{SIM}}(X,G) = 1 - cosine(Average(X), Average(G))$$

where $Average(X)$ is an embedding which is the mean over the word embeddings in sentence X.

- MatchingPyramid function(MP)[3]:

$$s_{mp}(X,G) = \mathrm{Matching-Pyramid}(X,G)$$

$$\mathrm{cons_{MP}}(X,G) = 1 - \frac{s_{mp}(X,G) - mn}{mx - mn}$$

where mn and mx are the min and max score of the score set(min-max normalization). We randomly select five negative generated sentences $\{GN_1, \dots, GN_5\}$. And the score set has six scores $\{s_{mp}(X,G), s_{mp}(X,GN_1), \dots, s_{mp}(X,GN_5)\}$

- BiLinear function(BL)[4]:

$$s_{bi}(X,G) = em(X) \times W \times em(G)$$

$$BL(X,G) = 1 - \frac{s_{bi}(X,G) - mn}{mx - mn}$$

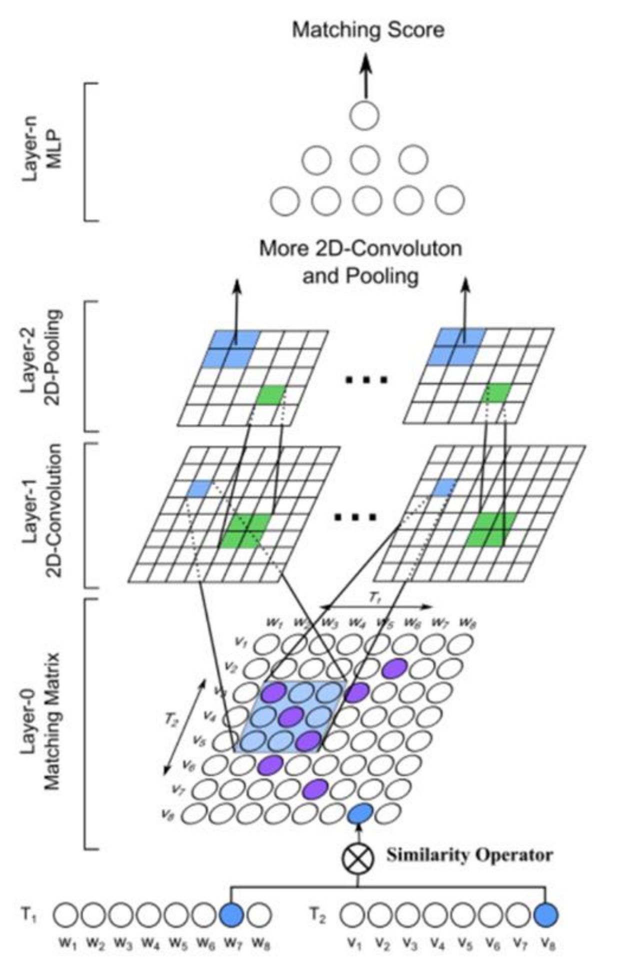
where $em(X)$ is the embedding of X with GRU encoder. W is a matrix of the transformation. ConsBL uses the same max-min normalization as MP does.

RETRIEVAL MODELS

We take use of **MatchPyramid**[3].

- Converting two 1D text representations of words within them to a typically 2D grid. Represent the input of text matching as a matching matrix M, with each element M_{ij} standing for the basic interaction, i.e. cosine similarity between word w_i and v_j.

- The body of MatchPyramid is a typical convolutional neural network, use the matching matrix mentioned below as input. The k-th kernel $w_{s,t}^{(1,k)}$ scans over the whole matching matrix M to generate a feature map $z_{i,j}^{(1,k)}$:



An overview of MatchPyramid on TextMatching

$$z_{i,j}^{(1,k)} = \sigma\left(\sum_{s=0}^{r_k-1}\sum_{t=0}^{r_k-1} w_{s,t}^{(1,k)} \cdot z_{i+s,j+t}^{(0)} + b^{(1,k)}\right)$$

where r_k denotes the size of the k-th kernel.

- A max-pooling is used to get a fixed length pattern vector. A two-layer DNN to produce the final matching score:

$$s = W_2\sigma(W_1 z + b_1) + b_2$$

GENERATION EXPERIMENTS

Table 1 The metric-based evaluation results generated from different models.

model	dist_uni	dist_bi	Average	Greedy	Extrema
Seq2Seq_att	0.004307	0.05082	0.5148	0.2815	0.3010
SIM	0.00424	0.05513	**0.5323**	0.2799	0.3088
MP	0.004294	0.05593	0.5267	0.2776	**0.3097**
BL	**0.00438**	**0.0587**	0.5322	**0.3048**	0.2877

Comparison results

- BL model has the best generations in metric-based evaluation. The distinct-bigram of BL model is 0.0587, which improves 15.5% compared with Seq2Seq _att.
- MP model is better. The distinct-bigram of MP model is 0.056, which improves 10.2% compared with Seq2Seq _att.
- The goal of SIM model is to optimized the Average measure, and it got the best evaluation on Average metric-based measure.

Conclusions

- Our constraint models have better results than Seq2Seq_att.
- The generated responses of our constraint models are more coherent to the post and the quality of generation has been improved.

RETRIEVAL EXPERIMENTS

Table 2 The metric results for different models on STC2.

model	MAP	NDCG@5	P@5
random	0.200	0.182	0.144
BM25	0.234	0.194	0.144
MatchPyramid	**0.484**	**0.404**	**0.276**

Experiment Settings

- We perform tokenization and discard the stop words, use word2vec to get word representation vectors with dimension 50. We use two different size of kernels in CNN and batch size as 200. We adopt Adam method with learning rate 0.1. We use BM25 and random results as baselines.

Conclusions

- Our model tries to find the hidden relationship between the query and the comment, which results in a better performance than the BM25 method.

REFERENCES

[1] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models [C]. Proceedings of NAACL-HLT. 2016: 110-119.
[2] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue[C]. EMNLP.2016.
[3] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng.Text matching as image recognition. AAAI2016 . Pages 2793–2799.
[4] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]. International Conference on Neural Information Processing Systems. Curran Associates Inc. 2013:926-934.