Xiang-zhou HUANG, Si-liang TANG, Yin ZHANG, Bao-gang WEI, 2020. Hybrid embedding and joint training of stacked encoder for opinion question machine reading comprehension. *Frontiers of Information Technology & Electronic Engineering*, 21(9):1346-1355. https://doi.org/10.1631/FITEE.1900571

Hybrid embedding and joint training of stacked encoder for opinion question machine reading comprehension

Key words: Machine reading comprehension; Neural networks; Joint training; Data augmentation

Corresponding author: Yin ZHANG
E-mail: yinzh@zju.edu.cn
ORCID: https://orcid.org/0000-0001-6986-4227

Motivation

1. During the competition AI Challenger 2018, external data are not allowed. The common methods which pre-train word- or character-level embedding independently do not work well with such a small-scale corpus.

2. As encoders to a long text, stacked LSTMs often give better results, but they are also difficult to train because of exploding and vanishing gradient problems.

3. The distribution of three labels in the training set is unbalanced and exerts a bad influence on model training.

Main idea

1. We combine POS tags to pre-train hybrid embedding, which contains more semantic information and can represent questions and passages better.

2. We jointly optimize all *K* multi-class cross-entropy losses corresponding to every stacked bi-directional LSTM in the passage encoding layer during the training process.

3. Irrelevancy of question and passage is used for data augmentation and new samples with label unidentified are generated. The new samples are added to the training set, and higher accuracy is achieved in experiments.

Method

1. We combine POS tags to pre-train hybrid embedding.

Algorithm 1 Corpus tagging

Input: segmented word list of a corpus ($W_{\mathrm{C}} = \{w_n\}_{n=1}^{N_{\mathrm{C}}}$), POS tags of a segmented corpus ($P_{\mathrm{C}} = \{p_n\}_{n=1}^{N_{\mathrm{C}}}$), and the minimum word-frequency (F_{m})

Output: hybrid tag for every segmented word in the corpus ($T_{\mathrm{C}} = \{t_n\}_{n=1}^{N_{\mathrm{C}}}$)

$D = \{\}$ // Initialize an empty dictionary
for w_n in W_{C} **do**
  if w_n not in D **then**
    $D[w_n] = 1$ // Save a new word-frequency
  else
    $D[w_n]$ += 1 // Count the word-frequency
  end if
end for
$T_{\mathrm{C}} = [\,]$
P_{C}.insert($\hat{R}$, 0)
P_{C}.insert($\hat{R}$, −1)
for w_n in W_{C} **do**
  if w_n in D and $D[w_n] >= F_{\mathrm{m}}$ **then**
    T_{C}.append($w_n \oplus p_n$)
  else
    T_{C}.append($p_{n-1} \oplus \hat{R} \oplus p_{n+1}$)
    $p_n = \hat{R}$
  end if
end for
Return T_{C}

Method (Cont'd)

2. We jointly optimize all *K* multi-class cross-entropy losses.

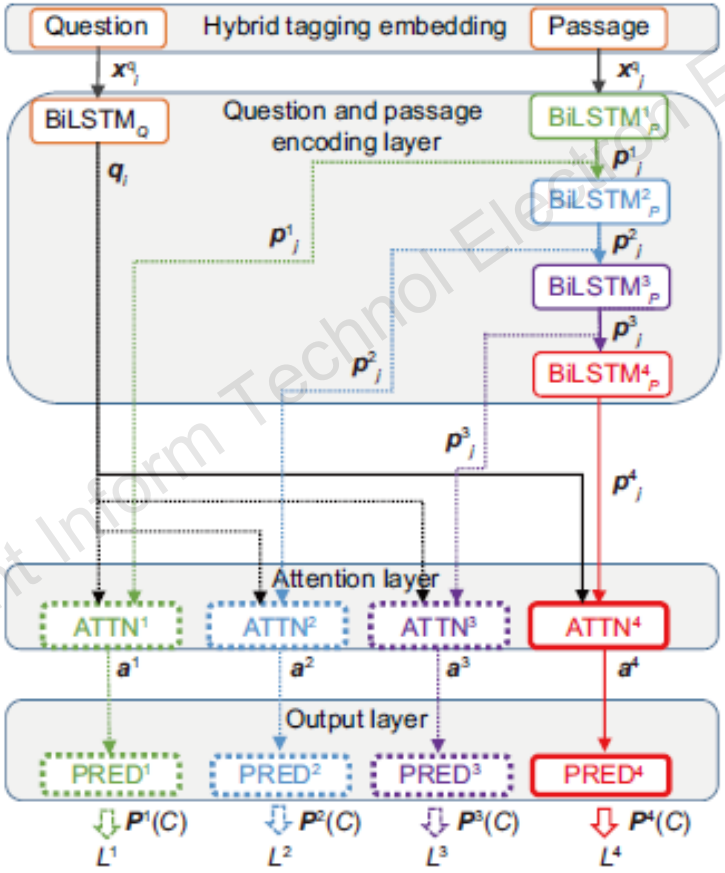

Fig. 2 Overview of the proposed approach based on neural networks

Method (Cont'd)

3. Irrelevancy of question and passage is used for data augmentation.

Question:
做礼拜能不能玩手机
Can I play on my cellphone at church?

Converted statement:
做礼拜能玩手机
I can play on my cellphone at church

Passage:
19%的美国人在去教堂做礼拜时玩手机。75%的美国人在任何时候手机都不会超出距离自己1.5米的范围。54%的美国人在床上玩手机，无论是睡觉前还是半夜醒来时。
19% of Americans play on cellphones while going to church. 75% of Americans keep their cellphones within 1.5 m from themselves. 54% of Americans play on cellphones in bed, no matter before sleeping or waking up in the midnight.

Label: true

{token$_q$}: {手机(cellphone)}

{token$_p$}: {手机(cellphone), 美国人(American)}

{token$_{qp}$}: {手机(cellphone)}

{s_{qp}}:
19%的美国人在去教堂做礼拜时玩手机。75%的美国人在任何时候手机都不会超出距离自己1.5米的范围。54%的美国人在床上玩手机
19% of Americans play on cellphones while going to church. 75% of Americans keep their cellphones within 1.5 m from themselves. 54% of Americans play on cellphones in bed

{s_{nqp}}:
无论是睡觉前还是半夜醒来时
no matter before sleeping or waking up in the midnight

Fig. 3 An example of data augmentation

Major results

Table 4 Accuracies of the proposed approach and competing systems in the dataset for the AIC2018 OQMRC task

Method	Accuracy (%)	
	Dev	Test
Official baseline (Tan et al., 2018)	69.52	69.90
AR (Hermann et al., 2015)	65.32	66.04
Match-LSTM (Wang SH and Jiang, 2016)	70.25	70.99
BIDAF (Seo et al., 2016)	72.30	72.56
R-NET (Wang W et al., 2017)	73.66	74.14
QANet (Yu et al., 2018)	61.37	62.11
BERT (Devlin et al., 2018)	70.65	70.99
Our approach	**76.35**	**77.52**

Dev: development. The best results are in bold

Major results (Cont'd)

Table 5 Ablation performance of the proposed approach

Method	Accuracy (%)	
	Dev	Test
Our approach	76.35	77.52
-JT	73.12 (−3.23)	73.96 (−3.56)
-JT+RC	75.28 (−1.07)	N/A
+RC	76.26 (−0.09)	77.33 (−0.19)
-HE	74.65 (−1.70)	76.03 (−1.49)
-HE+CE	74.77 (−1.58)	N/A
-HE+PE	74.55 (−1.80)	N/A
-HE+CE+PE	74.71 (−1.64)	N/A
+CE+PE	75.84 (−0.51)	76.97 (−0.55)
-DA	75.44 (−0.91)	N/A

-: exclude; +: include. JT: joint training; HE: hybrid embedding; DA: data augmentation; RC: residual connections; CE: character-level embedding; PE: POS tag embedding; Dev: development. N/A: not applicable

Conclusions

1. POS tags have been combined to enrich the semantic representation of questions and passages.

2. Extra attention and output layers have been introduced in the training process, and multiple losses have been jointly optimized to better update the parameters of networks.

3. To relieve the problem of data imbalance in the competition, a data augmentation strategy has been implemented to generate new samples.