Fine-grained Embedding for Reading Comprehension

Shanghai Jiao Tong University

Reporter: Zhuosheng Zhang

zhangzs@sjtu.edu.cn

2017.10.14

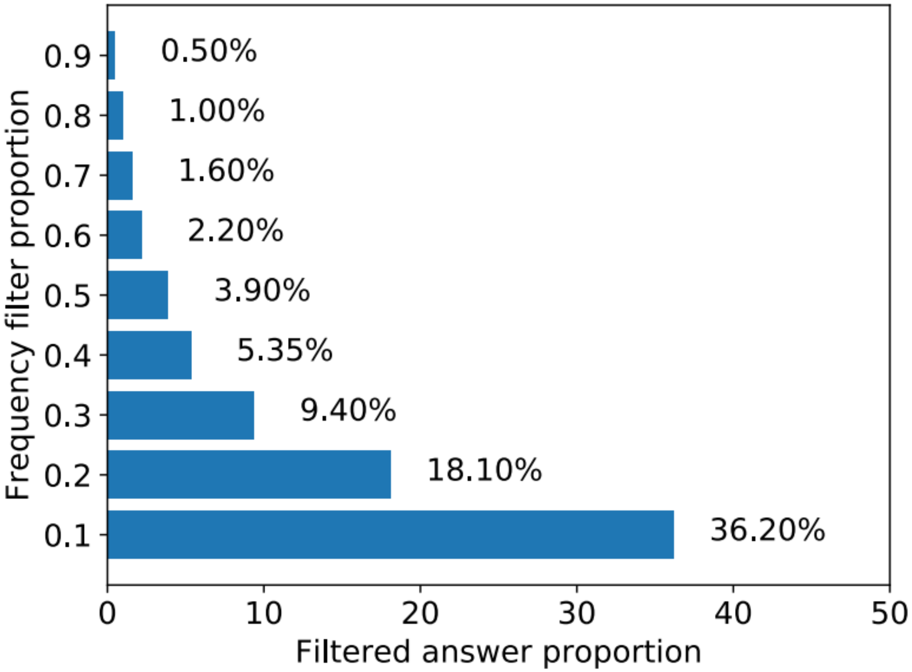
Task

The cloze-style task can be described as a triple ‹ *D*; *Q*; *A* ›, where *D* is a document (context), *Q* is a query over the contents of *D*, in which a word or phrase is replaced with a placeholder, and *A* is the answer to *Q*.

Document	1 早上，青蛙、小白兔、刺猬和大蚂蚁高高兴兴过桥去赶集。 2 不料，中午下了一场大暴雨，哗啦啦的河水把桥冲走了。 3 天快黑了，小白兔、刺猬和大蚂蚁都不会游泳。 4 过不了河，急得哭了。 5 这时，青蛙想，我可不能把朋友丢下，自己过河回家呀。 6 他一面劝大家不要着急，一面动脑筋。 7 嗬，有了！ 8 他说：“我有个朋女住在这儿，我去找他想想办法。 9 青蛙找到了他的朋友________，请求他说：“大家过不了河了，请帮个忙吧！ 10 鼹鼠说：“可以，请把大家领到我家里来吧。 11 鼹鼠把大家带到一个洞口，打开了电筒，让小白兔、刺猬、大蚂蚁和青蛙跟着他，“大家别害怕，一直朝前走。 12 走呀走呀，只听见上面“哗啦哗啦”的声音，象唱歌。 13 走着走着，突然，大家看见了天空，天上的月亮真亮呀。 14 小白兔回头一瞧，高兴极了：“哈，咱们过了河啦！ 15 唷，真了不起。 16 原来，鼹鼠在河底挖了一条很长的地道，从这头到那头。 17 青蛙、小白兔、刺猬和大蚂蚁是多么感激鼹鼠啊！ 18 第二天，青蛙、小白兔、刺猬和大蚂蚁带来很多很多同伴，扛着木头，抬着石头，要求鼹鼠让他们来把地道挖大些，修成河底大“桥”。 19 不久，他们就把鼹鼠家的地道，挖成了河底的一条大隧道，大家可以从河底过何，还能通车，真有劲哩!	1 In the morning, the frog, the little white rabbit, the hedgehog and the big ant happily crossed the bridge for the market. 2 Unexpectedly, a heavy rain fell at noon, and the water swept away the bridge. 3 It was going dark. The little white rabbit, hedgehog and big ant cannot swim. 4 Unable to cross the river, they were about to cry. 5 At that time, the frog made his mind that he could not leave his friend behind and went home alone. 6 Letting his friends take it easy, he thought and thought. 7 Well, there you go! 8 He said, “I have a friend who lives here, and I'll go and find him for help.” 9 The frog found his friend ________ and told him, “We cannot get across the river. Please give us a hand!” 10 The mole said, "That's fine, please bring them to my house." 11 The mole took everyone to a hole, turned on the flashlight and asked the little white rabbit, the hedgehog, the big ant and the frog to follow him, saying, "Don't be afraid, just go ahead." 12 They walked along, hearing the “walla-walla” sound, just like a song. 13 All of a sudden, everyone saw the sky, and the moon was really bright. 14 The little white rabbit looked back and rejoiced: “ha, the river crossed!”. 15 “Oh, really great.” 16 Originally, the mole dug a very long tunnel under the river, from one end to the other. 17 How grateful the frog, the little white rabbit, the hedgehog and the big ant felt to the mole! 18 The next day, the frog, the little white rabbit, the hedgehog, and the big ant with a lot of his fellows, took woods and stones. They asked the mole to dig tunnels bigger, and build a great bridge under the river. 19 It was not long before they dug a big tunnel under the river, and they could pass the river from the bottom of the river, and it could be open to traffic. It is amazing!
Query	青蛙找到了他的朋友________，请求他说：“大家过不了河了，请帮个忙吧！”	The frog found his friend ________ and told him, “We cannot get across the river. Please give us a hand!”
Answer	鼹鼠	the mole

OOV issues

Reading comprehension systems usually suffer from out-of-vocabulary (OOV) word issues, especially when the ground-truth answers contain rare words or name entities, which are hardly fully recorded in the vocabulary.



There are over 13,000 characters in Chinese while there are only 26 letters in English without regard to punctuation marks.

If a reading comprehension system can not effectively manage the OOV issues, the performance will not be semantically accurate for the task.

Two levels of embedding

Word-level Embedding

青蛙|和|小白兔|去|赶集

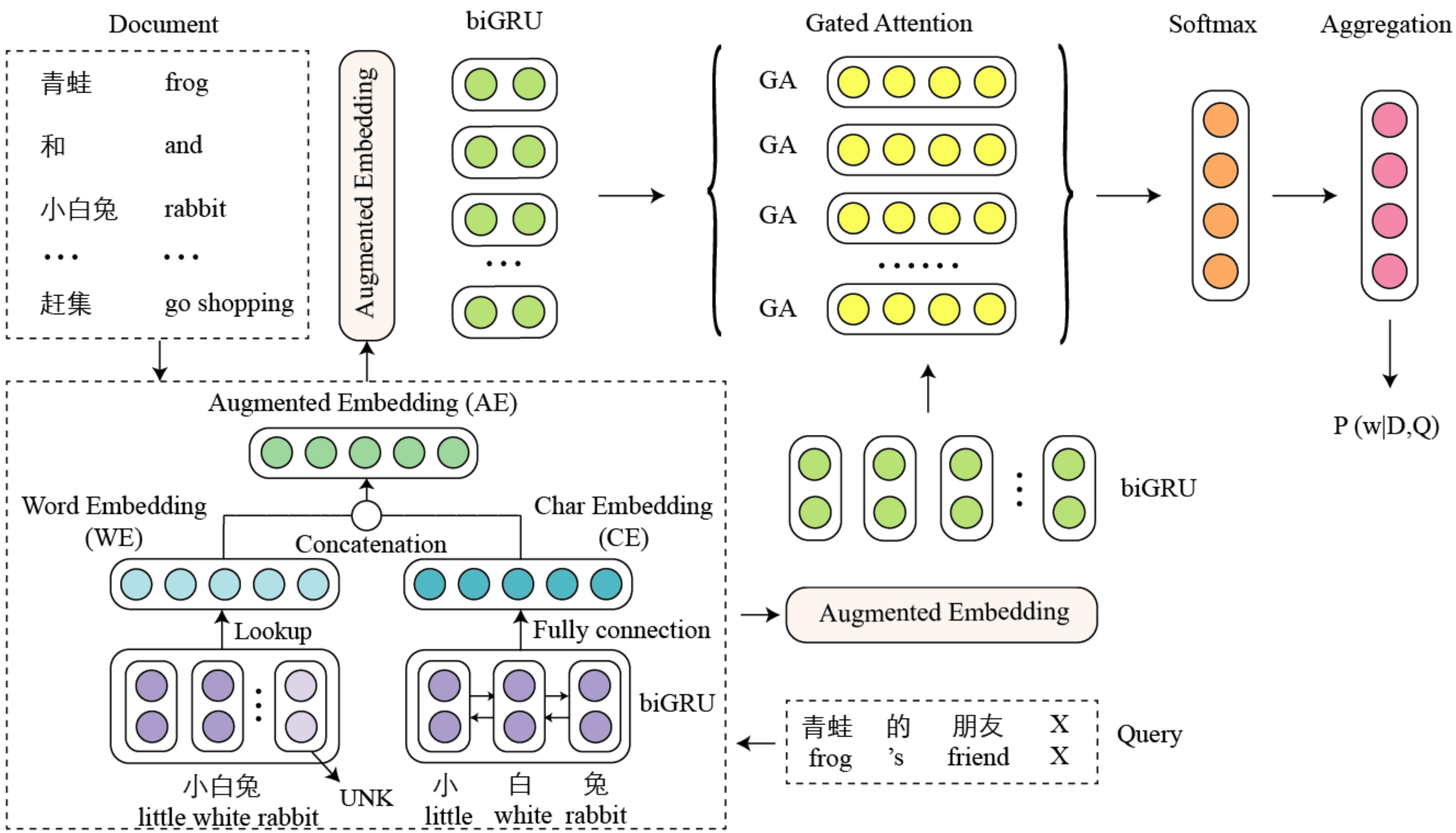
Character-level Embedding

青|蛙|和|小|白|兔|去|赶|集

- Intuitively, word-level representation is good at catching global context and dependency relationships between words. However, rare words are often expressed poorly due to data sparsity.
- Character embedding are more expressive to model sub-word morphologies, which is beneficial to deal with rare words. However, quite a lot of Chinese words, like "**吉(auspicious)普(ordinary)**" (jeep) are not semantically character-level compositional at all.
- Using extra features, such as named entity recognition (NER) and part-of-speech (POS) tagging will result in tremendous computational complexity.

Framework

- Given the triple ‹ $D; Q; A$ ›, the system will be built in the following steps.



Filtered Lookup

Trainable Embedding

Motivation: insufficient training for UNK words

Technique:

- Sort the dictionary according to the word frequency from high to low.
- A frequency filter ratio γ is set to filter out the low-frequency words (rare words) from the lookup table.
- For example, if γ is 0.9, then the last 10% low-frequency words will be mapped into UNK words.

的
了
一
小
我
说
在
是
不
你
着
他
……
药膏
洪武私访
彩虹曲
牢合·乔治
攻坚
厅长

High-frequency words (90%)

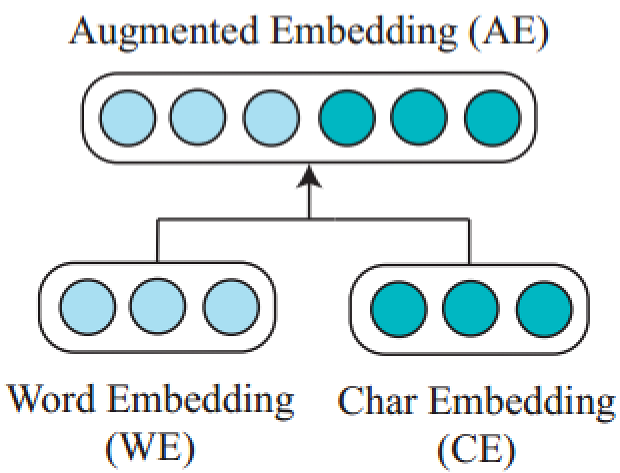
γ = 0.9

low-frequency words (10%)

Fine-grained Embedding

- Word embedding WE(w) is indexed from word lookup table

- Characters of each word are successively fed to the forward GRU and backward GRU. The output for each input is the concatenation of the two vectors from both directions: $\overleftrightarrow{h_t} = \overrightarrow{h_t} \parallel \overleftarrow{h_t}$



- The augmented embedding (AE) is given by concatenating the word embedding and character-level representation. $AE(w) = WE(w) \,||\, CE(w)$

Gated-attention Leaning (Dhingra et al. 2017)

- Contextual representations of the document and query

$$H_q = \text{BiGRU}(Q)$$
$$H_d = \text{BiGRU}(D)$$

- Gated-attention

$$\alpha_i = softmax(H_q^\top d_i)$$
$$\beta_i = Q\alpha_i$$
$$x_i = d_i \odot \beta_i$$

- Probability of each candidate word as being the answer

$$p = softmax((q_t)^\top H_D)$$
$$P(w|D,Q) \propto \sum_{i \in I(w,D)} p_i$$

- The predicted answer

$$A^* = \text{argmax}_{w \in C} P(w|D,Q)$$

Dataset and hyper-parameters

		Cloze Track		User Query Track	
	Train	Validation	Test	Validation	Test
# Query	354,295	2,000	3,000	2,000	3,000
Max # tokens in docs	486	481	484	481	486
Max # tokens in query	184	72	106	21	29
Avg # tokens in docs	324	321	307	310	290
Avg # tokens in query	27	19	23	8	8
Vocabulary			94,352		

Word embedding： pre-trained by word2vec toolkit on Wikipedia corpus

Optimization: stochastic gradient descent with ADAM updates for optimization

Batch size : 32

Learning rate: 0.001 (halved every epoch)

Word embedding size	200
Character embedding size	100
Hidden unit number	128
Dropout	0.5
Default frequency filter proportion	0.9

Result

填空类问题（Cloze-style Question）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	**6ESTATES PTE LTD**	**多系统**	81.85%	81.90%
2	**上海交通大学仿脑计算与机器智能研究中心自然语言组** Shanghai Jiao Tong University (SJTU BCMI-NLP)	**多系统**	78.35%	80.67%
3	**南京云思创智信息科技有限公司**	**多系统**	79.20%	80.27%

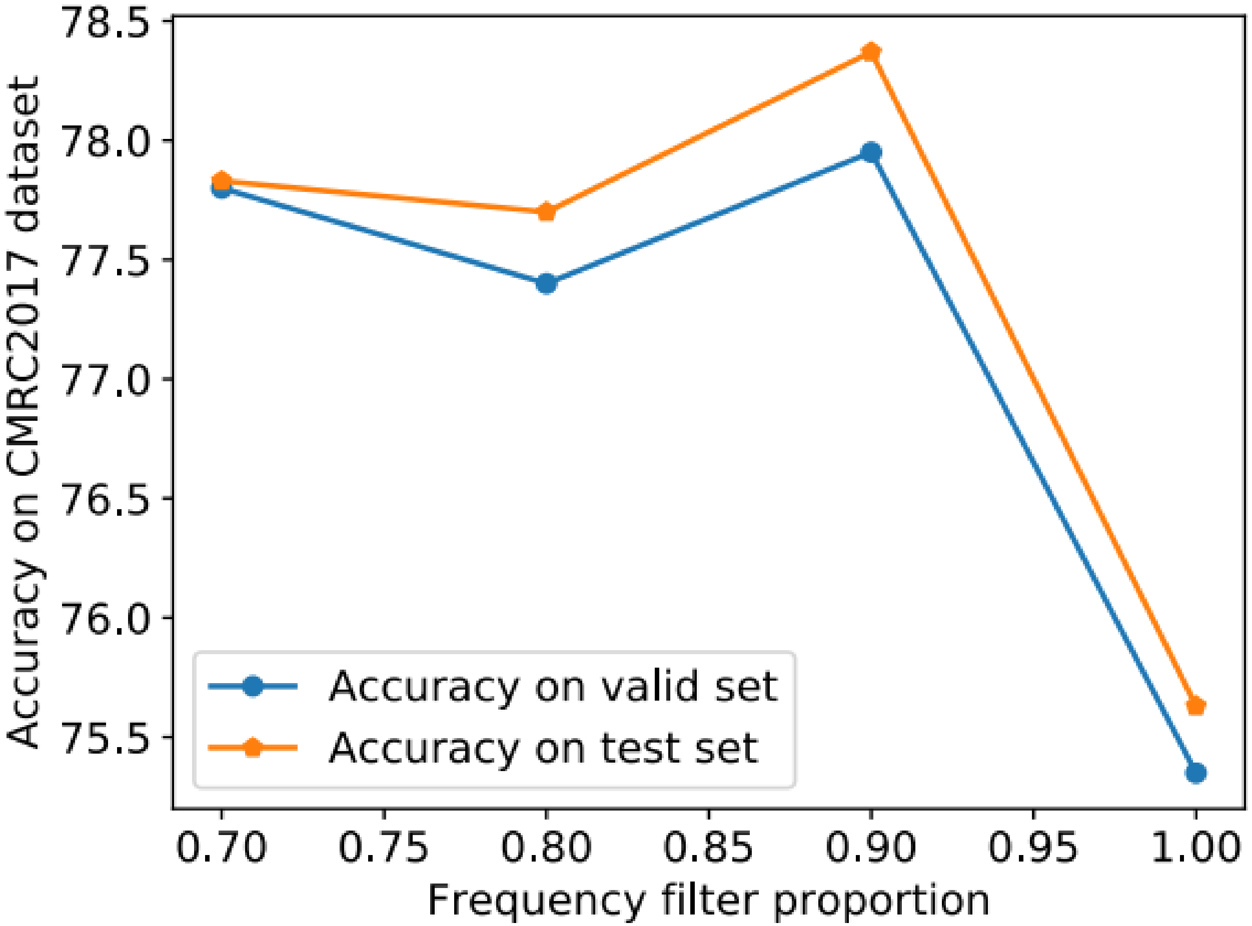
用户提问类问题（User-Query Question）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	**华东师范大学** East China Normal University (ECNU)	**多系统**	90.45%	69.53%
2	**山西大学三队** Shanxi University (SXU-3)	单系统	47.80%	49.07%
3	**郑州大学** Zhengzhou University (ZZU)	单系统	31.10%	32.53%

最佳单系统（Best Single System）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	**上海交通大学仿脑计算与机器智能研究中心自然语言组** Shanghai Jiao Tong University (SJTU BCMI-NLP)	单系统	76.15%	77.73%

Analysis



78.5
78.0
77.5
77.0
76.5
76.0
75.5
Accuracy on CMRC2017 dataset
Accuracy on valid set
Accuracy on test set
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Frequency filter proportion

Thanks!
Q&A