Subword-augmented Embedding for Cloze Reading Comprehension

Zhuosheng Zhang[1,2,*], **Yafang Huang**[1,2,*], **Hai Zhao**[1,2,†]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
{zhangzs, huangyafang}@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Task

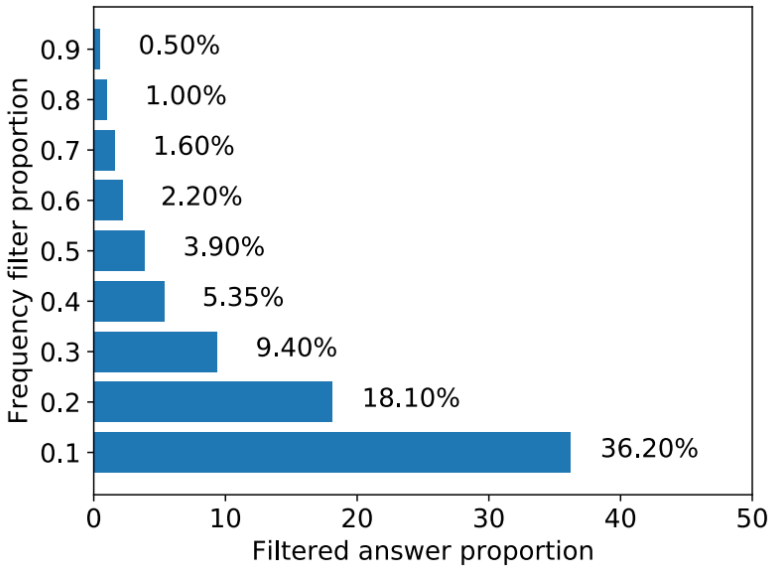
This work focuses on a cloze-style reading comprehension task over fairy stories, which is highly challenging due to diverse semantic patterns with personified expressions and reference.

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 嘀，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 嘻，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，杠着木头，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

Representation challenges

- Representation difficulty and computational complexity due to the large vocabulary and data sparsity.
- 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 common levels of embedding

Word-level Embedding

青蛙|和|小白兔|去|赶集

Character-level Embedding

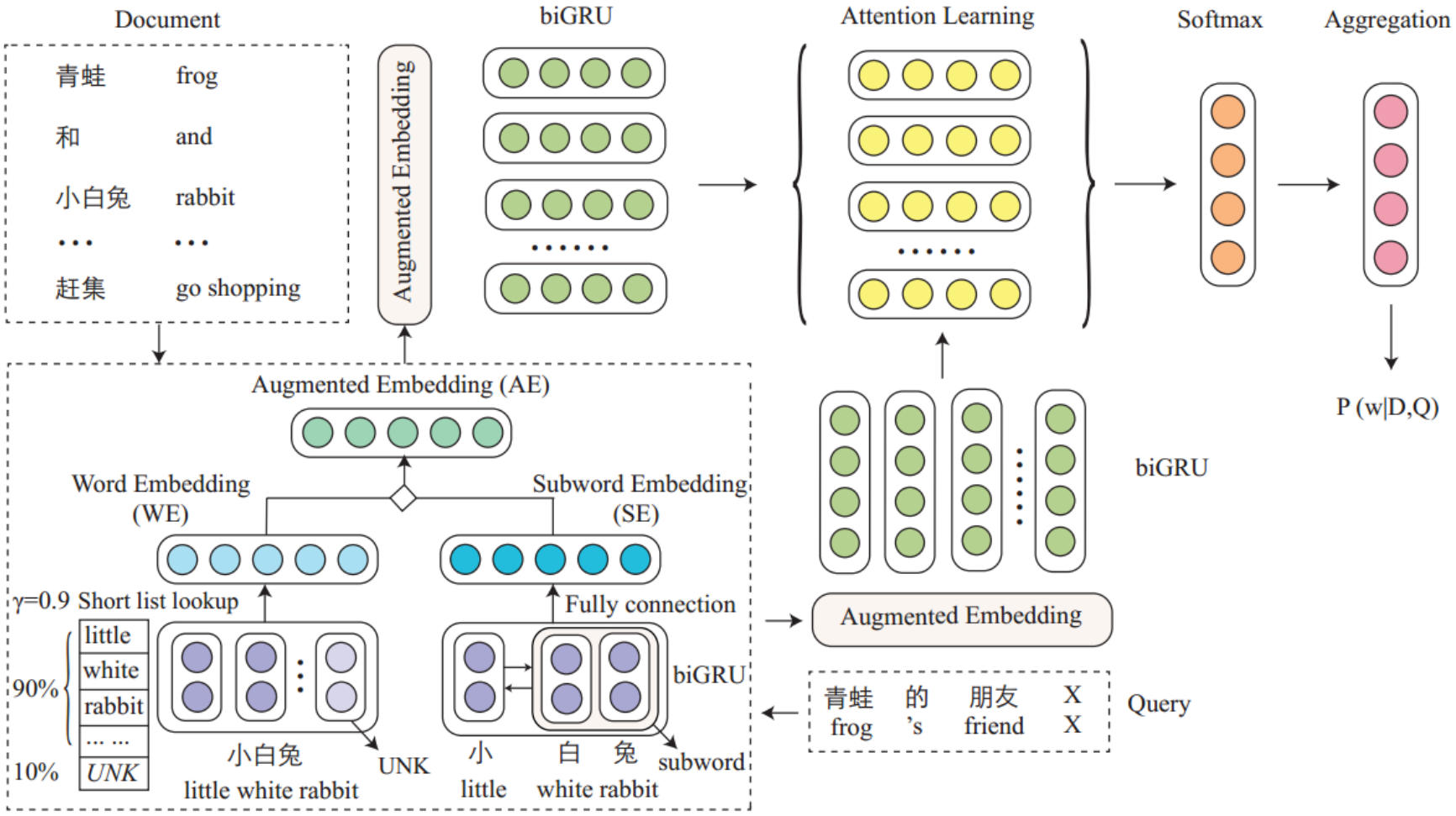
青|蛙|和|小|白|兔|去|赶|集

- **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, the minimal meaningful unit below word usually is not character, which motivates researchers to explore the potential unit (**subword**) between character and word to model sub-word morphologies or lexical semantics.

Word	Subword
indispensability	in disp ens ability
intercontinentalexchange	inter contin ent al ex change
playgrounds	play ground s
大花猫	大 花猫
一步一个脚印	一步 一个 脚印

Framework

- Given the triple < *D; Q; A* >, the system will be built in the following steps.



BPE Subword Segmentation

Word in most languages usually can be split into meaningful **subword units** despite of the writing form.

For example, "*indispensable*" could be split into < *in; disp; ens; able* >.

The generalized framework:

Firstly, all the input sequences (strings) are tokenized into a sequence of single-character subwords, then we repeat:

1. Count all **bigrams** under the current segmentation status of all sequences.
2. Find the bigram with the **highest frequency** and merge them in all the sequences. Note the segmentation status is updating now.
3. If the merging times do not reach the specified number, go back to 1, otherwise the algorithm ends.

Subword-augmented Word Embedding

An augmented embedding (***AE***) is to straightforwardly integrate word embedding ***WE***(*w*) and subword embedding ***SE***(*w*) for a given word *w*.

$$AE(w) = WE(w) \diamond SE(w)$$

In this work, we investigate concatenation (*concat*), element-wise summation (*sum*) and element-wise multiplication (*mul*).

The subword embedding *SE*(*w*) is generated by taking the final outputs of a bidirectional gated recurrent unit (GRU)

$$SE(w) = W \overleftrightarrow{h_t} + b$$

Short list lookup

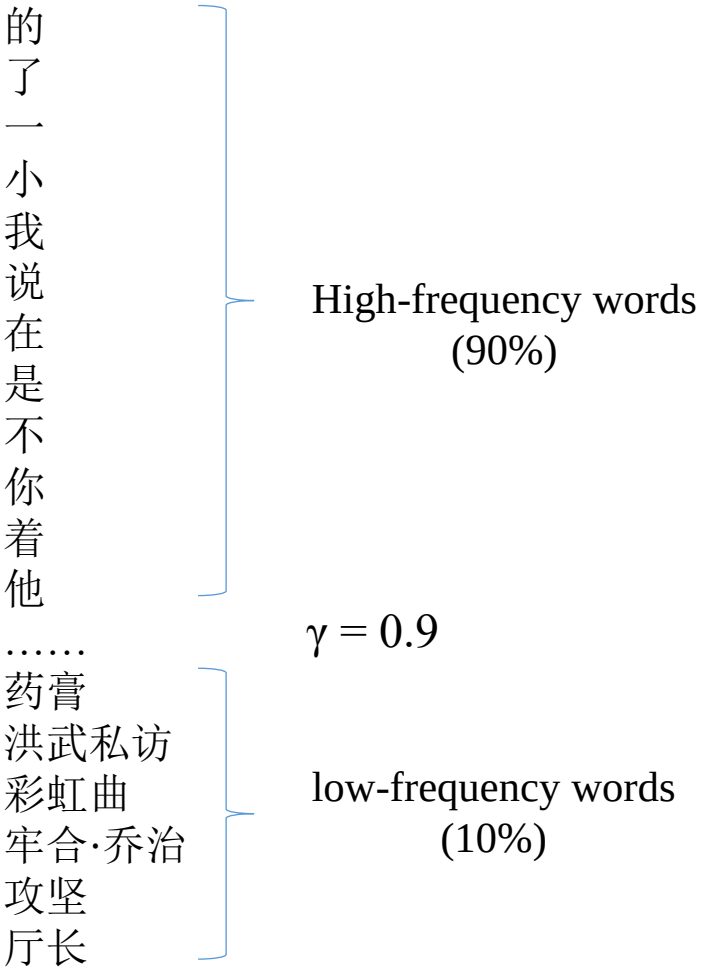
Trainable Embedding

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.
- Thus, *AE*(*w*) can be rewritten as

$$AE(w) = \begin{cases} WE(w) \diamond SE(w) & \text{if } w \in H \\ UNK \diamond SE(w) & \text{otherwise} \end{cases}$$



Attention Module

- 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

	CMRC-2017			PD			CFT
	Train	Valid	Test	Train	Valid	Test	human
# Query	354,295	2,000	3,000	870,710	3,000	3,000	1,953
Max # words in docs	486	481	484	618	536	634	414
Max # words in query	184	72	106	502	153	265	92
Avg # words in docs	324	321	307	379	425	410	153
Avg # words in query	27	19	23	38	38	41	20
# Vocabulary	94,352	21,821	38,704	248,160	536	634	414

- Three Chinese Machine Reading Comprehension datasets, namely CMRC-2017, People's Daily (PD) and Children Fairy Tales (CFT).
- We also use the Children's Book Test (CBT) dataset (Hill et al., 2015) to test the generalization ability in multi-lingual case.

Main results

- Our SAW Reader (***mul***) outperforms all other single models
- ***mul*** might be more informative than ***concat*** and ***sum*** operations

Model	CMRC-2017	
	Valid	Test
Random Guess †	1.65	1.67
Top Frequency †	14.85	14.07
AS Reader †	69.75	71.23
GA Reader	72.90	74.10
SJTU BCMI-NLP †	76.15	77.73
6ESTATES PTE LTD †	75.85	74.73
Xinktech †	77.15	77.53
Ludong University †	74.75	75.07
ECNU †	77.95	77.40
WHU †	78.20	76.53
SAW Reader	**78.95**	**78.80**

Model	Operation	CMRC-2017	
		Valid	Test
Word + Char	concat	74.80	75.13
	sum	75.40	75.53
	mul	77.80	77.93
Word + BPE	concat	75.95	76.43
	sum	76.20	75.83
	mul	**78.95**	**78.80**

Table 3: Case study on CMRC-2017.

Model	PD		CFT
	Valid	Test	Test-human
AS Reader	64.1	67.2	33.1
GA Reader	67.2	69.0	36.9
CAS Reader	65.2	68.1	35.0
SAW Reader	**72.8**	**75.1**	**43.8**

Accuracy on CBT dataset

Our model outperforms most of the previously public works.

Model	CBT-NE		CBT-CN	
	Valid	Test	Valid	Test
Human ‡	-	81.6	-	81.6
LSTMs ‡	51.2	41.8	62.6	56.0
MemNets ‡	70.4	66.6	64.2	63.0
AS Reader ‡	73.8	68.6	68.8	63.4
Iterative Attentive Reader ‡	75.2	68.2	72.1	69.2
EpiReader ‡	75.3	69.7	71.5	67.4
AoA Reader ‡	77.8	72.0	72.2	69.4
NSE ‡	78.2	73.2	74.3	71.9
FG Reader ‡	**79.1**	**75.0**	**75.3**	**72.0**
GA Reader ‡	76.8	72.5	73.1	69.6
SAW Reader	78.5	74.9	75.0	71.6

Analysis

- When the vocabulary size is **1*k*** and γ = **0.9**, the models could obtain the best performance.
- For a task like reading comprehension the subwords, being a highly flexible grained representation between character and word, tends to be more like **characters** instead of words.
- The **balance** between word and character is quite critical and an appropriate grain of character-word segmentation could essentially improve the word representation

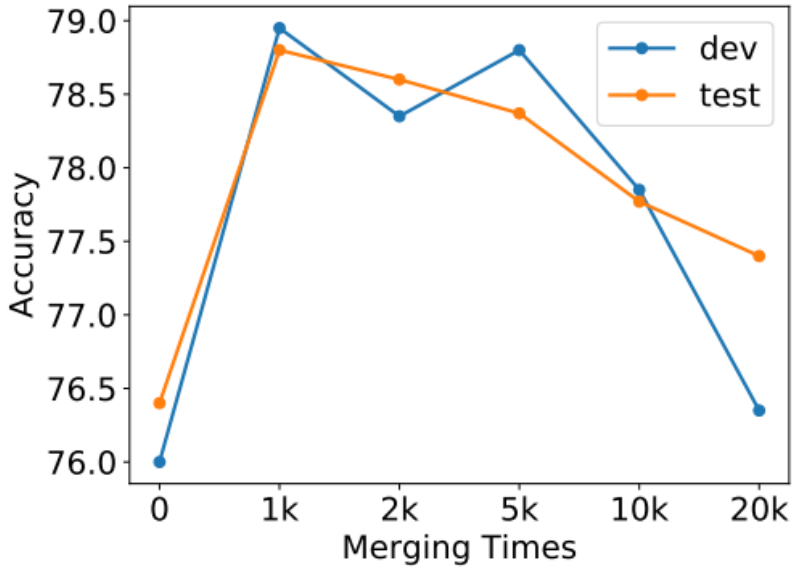


Figure 2: Case study of the subword vocabulary size of BPE.

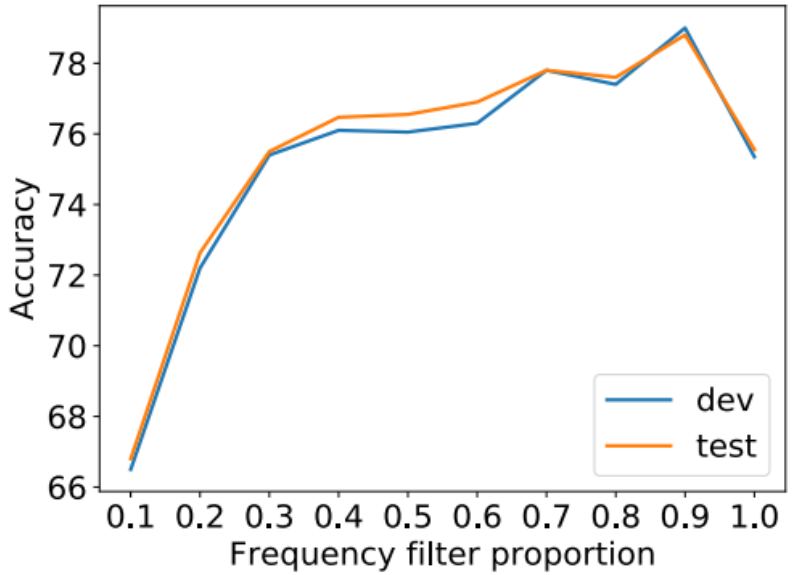
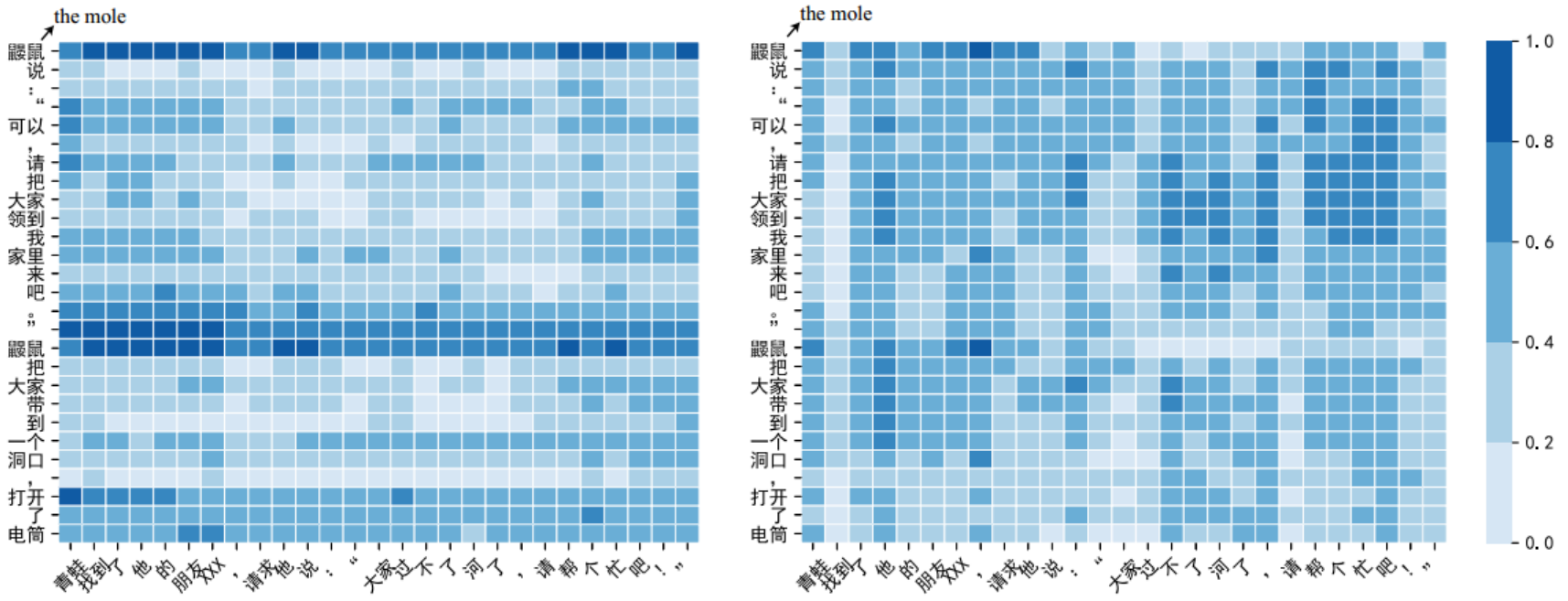


Figure 3: Quantitative study on the influence of the short list.

Subword-Augmented Representations

- In CMRC-2017, we observe questions with OOV answers (denoted as “OOV questions”) account for **17.22%** in the error results of the best Word + Char embedding based model.
- With BPE subword embedding, **12.17%** of these “OOV questions” could be correctly answered.
- This shows the subword representations could be essentially useful for modeling rare and unseen words.



(a) Embedding of Document and query

(b) Final document and query representation

Doc (extract): The mole said, "That's fine, please bring them to my house." 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."
Query: The frog found his friend _____ and told him, We cannot get across the river. Please give us a hand!

Conclusion

- This paper presents an effective neural architecture, called **subword-augmented word embedding** to enhance the model performance for the cloze-style reading comprehension task.
- The proposed SAW Reader uses subword embedding to enhance the word representation and limit the word frequency spectrum to train rare words efficiently.
- With the help of the **short list**, the model size will also be reduced together with training speedup.
- Giving state-of-the-art performance on multiple benchmarks, the proposed reader has been proved effective for learning joint representation at both word and subword level and alleviating OOV difficulties.

Thanks!

Q&A