Fine-grained Embedding for Reading Comprehension

Shanghai Jiao Tong University

Reporter: Zhuosheng Zhang

zhangzs@sjtu.edu.cn

2017.10.14
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 \( Q \).
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 cannot 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.
• Given the triple $\langle D; Q; A \rangle$, 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 $\gamma$ is set to filter out the low-frequency words (rare words) from the lookup table.

• For example, if $\gamma$ is 0.9, then the last 10% low-frequency words will be mapped into UNK words.

\[
\begin{align*}
\text{High-frequency words (90\%)} \\
\gamma = 0.9 \\
\text{low-frequency words (10\%)}
\end{align*}
\]
The augmented embedding (AE) is given by concatenating the word embedding and character-level representation. 

\[ AE(w) = WE(w) \| CE(w) \]

- 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: 
  \[ \vec{h}_t = \vec{h}_t \| \vec{h}_t \]
- The augmented embedding (AE) is given by concatenating the word embedding and character-level representation. 

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
  \[ \alpha_i = \text{softmax}(H_q^T d_i) \]
  \[ \beta_i = Q \alpha_i \]
  \[ x_i = d_i \odot \beta_i \]

- Probability of each candidate word as being the answer
  \[ p = \text{softmax}((q_t)^T H_D) \]
  \[ P(w|D, Q) \propto \sum_{i \in I(w, D)} p_i \]

- The predicted answer
  \[ A^* = \arg\max_{w \in C} P(w|D, Q) \]
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)
### 填空类问题（Cloze-style Question）

<table>
<thead>
<tr>
<th>最终排名</th>
<th>参赛单位</th>
<th>单/多系统</th>
<th>开发集准确率</th>
<th>测试集准确率</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6ESTATES PTE LTD</td>
<td>多系统</td>
<td>81.85%</td>
<td>81.90%</td>
</tr>
<tr>
<td>2</td>
<td>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)</td>
<td>多系统</td>
<td>78.35%</td>
<td>80.67%</td>
</tr>
<tr>
<td>3</td>
<td>南京云思创智信息科技有限公司</td>
<td>多系统</td>
<td>79.20%</td>
<td>80.27%</td>
</tr>
</tbody>
</table>

### 用户提问类问题（User-Query Question）

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

### 最佳单系统（Best Single System）

<table>
<thead>
<tr>
<th>最终排名</th>
<th>参赛单位</th>
<th>单/多系统</th>
<th>开发集准确率</th>
<th>测试集准确率</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>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)</td>
<td>单系统</td>
<td>76.15%</td>
<td>77.73%</td>
</tr>
</tbody>
</table>
Analysis

![Graph showing accuracy on CMRC2017 dataset vs. frequency filter proportion. The blue line represents accuracy on the valid set, and the orange line represents accuracy on the test set.](image)
Thanks!

Q&A