Enhancing Embedding via Two-level Features for Machine Reading Comprehension

Shuai Wang, Hui Song, Bo Xu, Hongkuan Zhang
School of Computer Science and Technology, Donghua University, Shanghai, 201620, China

Contribution
- Through the algorithm analysis and case study of MRC case, we confirm the issues of existing models and propose a two-level approach for integrating token features and grammatical structure of taken-qna into sentence encoding.
- For token-level, we select the valid features from the candidate for different tokens according to the actual context, sentential vector is added. And for sentence level, we transform the dependency of sentence into NT-Matrix, and then integrate the information.
- We combine different features through to be introduced to conduct experiments, find that the impact of features varies on different models, therefore, we propose two level concatenation of MAC and RAN, get related models SQA-D1 base line, TIE-BERT base as baseline methods. Our model achieves 53.65% in SQA-D1 dataset and 66.37% in RAN-D1 dataset on the best result of previous methods, better than the current state-of-the-art models.

Method
- **Task**: Token extraction of MRC task, which predicts the answer by locating the start and end position in a context.
- **Input**: Given a question $q$, a document $d$, and a context $c$.
- **Output**: Predict the start and end position.

Modeling
- **Encoder**: Token-level encoder based on BERT, and enhance the BERT embedding two levels; token-level and sentence-level.
- **Token-level module**
  - **Input**: A sequence of tokens $[t_1, t_2, ..., t_n]$.
  - **Output**: A sequence of token embeddings $[e_1, e_2, ..., e_n]$.
  - **Operation**: Apply a pre-trained Transformer to encode each token.
  - **Integration**: Combine token embeddings using self-attention.

- **Sentence-level module**
  - **Input**: A sequence of token embeddings $[e_1, e_2, ..., e_n]$.
  - **Output**: A sequence of sentence embeddings $[s_1, s_2, ..., s_m]$.
  - **Operation**: Apply another pre-trained Transformer to encode each sentence.
  - **Integration**: Combine sentence embeddings using self-attention.

Datasets
- **Source**: 24,000 samples with a large portion of queries requiring common sense reasoning.
- **Questions are in the form of multiple-sentence questions.**
- **The 11 types of samples are summarized in the following table.

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>10,000</td>
</tr>
<tr>
<td>TIE-BERT based</td>
<td>2,000</td>
</tr>
<tr>
<td>RAN-D1</td>
<td>2,000</td>
</tr>
<tr>
<td>SQA-D1</td>
<td>2,000</td>
</tr>
<tr>
<td>Other</td>
<td>4,000</td>
</tr>
</tbody>
</table>

- **ReCoD**
  - 3,705 samples with a large portion of queries requiring common sense reasoning.
  - Questions are in the form of multiple-sentence questions.

Data Preprocessing
- **Standard**: MRC task usually needs: value and Extract Match (EM) for comprehensive evaluation.
  - **EM evaluation**: The degree of (statement of fact for prediction) and the correct answer.
  - **RAN-D1**: The number of the document that the number of correct answer character.
  - **SQA-D1**: The number of the number of overlap characters to the correct answer characters.

Model Evaluation
- **Environment**: Python 3.6, NVIDIA GeForce GTX 1080TI, Ubuntu 16.04.
- **Setup**: The trained model is 12 layers with 768 hidden layer dimension and 768 token-level feature is 20.

Result Analysis
- **General trend**: Compared TIE-BERT with three models DocQA, Google BERT-base and KT-NET (based on BERT base) in our environment.
  - **On SQA-D1 1-TIE-BERT gets an score of 53.65%**.
  - **On ReCoD, TIE-BERT achieves 66.37%**.

Conclusion
- **Future direction**: This paper proposes a new method of enhancing language representation through token-level and sentence-level features. We can select appropriate features as external information according to the characteristics of the dataset or the application background to improve the model performance.
- **Sentence-level dependency relation improves the performance of the model on both datasets, which also confirms our previous conjecture that the pre-trained language model does not obtain the sentence dependency as well.
- **Experiments have shown that our method is better than the current state-of-the-art models.**