Introduction

• Question answering (QA) is a classic and challenging machine comprehension task in the field of natural language processing.
• QA tests an NLP model’s ability to extract relevant, synthesize information from long pieces of text, and capture correlation between different passages. QA algorithms therefore have been applied to tasks as diverse as text summarization, code generation, and named entity recognition.

The Bi-directional Attention Flow (BIDAF) [3] model is a well-known QA architecture. In this project, I aimed to improve the performance of a BIDAF baseline by implementing a character level embedding layer, a self-attention layer, and an answer pointer output layer.

Problem Setup: SQuAD 2.0

I used the SQuAD 2.0 dataset [2] to train and evaluate all models. Each example in the dataset contains a “Context” or “Passage” and a “Question” or “Query.” A QA model must answer the question with a contiguous sequence of tokens found in the passage. Roughly half of the questions cannot be answered using the given passage. Correctly, to answer a question, the model outputs start and end indices that define a span within the context. If a question cannot be answered, the model outputs [-1, -1].

Example

Passage: European Union law is a body of treaties and legislation, such as Regulations and Directives, which have direct effect or indirect effect on the laws of European Union member states. The three sources of European Union law are primary law, secondary law and supplementary law. The main sources of primary law are the Treaties establishing the European Union. Secondary sources include regulations and directives which are based on the Treaties. The legislation of the European Union is principally composed of the European Parliament and the Council of the European Union, which under the Treaties, establish secondary law to pursue the objectives set out in the Treaties.

Question: What are the main sources of primary law?

Answer: Treaties establishing the European Union

Question: What is the last source of European Union law?

Answer: Cannot be answered

Dataset split:

12,914 examples in the train set, 6,578 in the dev set, and 5,913 in the test set.

Evaluation metrics:

Models are primarily evaluated based on their F1 scores. The Exact Match (EM) scores are also reported.

Methods

Three modifications were introduced to the baseline BIDAF model:

• The Character Embedding layer [3] uses a 1D CNN to process pre-trained character embeddings and generates a feature vector for each word.

• Inspired by Bi-Net [1], the Self-Attention layer is a residual layer that uses gated multiplicative attention to capture contextual relations within the passage. The attention outputs are concatenated with the inputs passed through an element-wise multiplicative gate, and normalized by a bi-directional GELU.

• The Answer Pointer layer [4] replaces the original BIDAF output layer and produces end point predictions conditioned on the start point prediction.

Model Architecture

The model architecture consists of the following components:

- **Modeling:** Input passage and query embeddings, multi-layer LSTM. Embeddings are obtained by the Highway Encoder.
- **Self-Attention:** Character level embedding layer.
- **Bi-Directional Attention:** Reduces the output dimension of the hidden state.
- **Answer Pointer:** Produces end point predictions conditioned on the start point prediction.

Training Details

Unless otherwise noted, all models were trained with the following hyperparameters:

- batch size = 64
- learning rate = 0.05
- dropout probability = 0.2
- L2 regularization weight = 0
- RNN hidden size = 100
- number of epochs = 30

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>dev set</th>
<th>test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM</td>
<td>F1</td>
</tr>
<tr>
<td>baseline</td>
<td>57.25</td>
<td>60.98</td>
</tr>
<tr>
<td>baseline + char embedding</td>
<td>60.36</td>
<td>63.48</td>
</tr>
<tr>
<td>baseline + self-attention</td>
<td>59.64</td>
<td>63.07</td>
</tr>
<tr>
<td>baseline + answer pointer</td>
<td>54.99</td>
<td>60.55</td>
</tr>
<tr>
<td>baseline + char embedding + self-attention</td>
<td>62.71</td>
<td>65.90</td>
</tr>
<tr>
<td>baseline + char embedding + self-attention + dropout = 0.1</td>
<td>64.36</td>
<td>67.40</td>
</tr>
</tbody>
</table>

References