Improving BiDAF: Character Embedding, Self-Attention, and Answer Pointer Output

Edward Gao
CS224N Final Project

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 word meaning, 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 **Bidirectional 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. Concretely, 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 legislature of the European Union is principally composed of the European Parliament and the Council of the European Union, which under the Treaties may establish secondary law to pursue the objective 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)

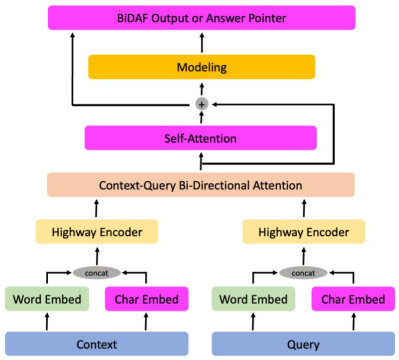
Dataset split:
129,941 examples in the train set, 6,078 in the dev set, and 5,915 in the test set.
Evaluation metrics:
Models are primarily evaluated based on their F1 scores. The Exact Match (EM) scores are also reported.

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 R-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 processed by a bi-directional GRU.
- The **Answer Pointer** layer [4] replaces the original BiDAF output layer and produces end point predictions conditioned on the start point prediction.

Model Architecture



Modified layers are highlighted in magenta.

Training Details

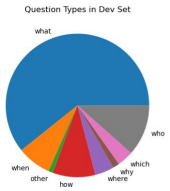
Unless otherwise noted, all models were trained with the following hyperparameters:

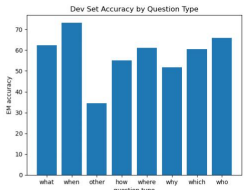
- batch size = 64
- learning rate = 0.5
- dropout probability = 0.2
- L2 regularization weight = 0
- RNN hidden size = 100
- number of epochs = 30

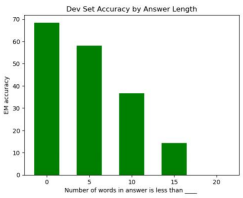
Results

Model	dev set EM	dev set F1	test set EM	test set F1
baseline	57.25	60.58	57.13	60.91
baseline + char embedding	60.36	63.48	-	-
baseline + self-attention	59.84	63.07	-	-
baseline + answer pointer	56.98	60.55	-	-
baseline + char embedding + self-attention	62.71	65.93	60.78	64.24
baseline + char embedding + self-attention + dropout = 0.1	64.36	67.40	62.30	65.84

Analysis







Conclusions

- Character embedding and self-attention improved the performance of the BiDAF model, but the answer pointer layer did not.
- Model performance varied based on question type. "When" questions were the easiest to answer, while "why" and "other" questions were the hardest.
- The model was significantly better at producing short answers (including no answers) than longer ones.

References

[1] Natural Language Computing Group.
R-net: Machine reading comprehension with self-matching networks.
May 2017.

[2] Pranav Rajpurkar, Robin Jia, and Percy Liang.
Know what you don't know: Unanswerable questions for SQuAD.
In *Association for Computational Linguistics (ACL)*, 2018.

[3] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi.
Bidirectional attention flow for machine comprehension.
CoRR, abs/1611.01603, 2016.

[4] Shuohang Wang and Jing Jiang.
Machine comprehension using match-lstm and answer pointer.
CoRR, abs/1608.07905, 2016.