Exploring the effects of semantic tag augmentations for SQuAD

Jonathan Tseng [1]   Narvin Phouksouvath [1]

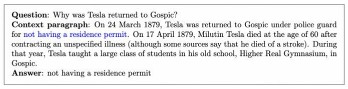
[1]Computer Science, Stanford

Problem

Question-Answering is a broadly researched NLP problem which has seen many contributions. Many publications that approach the QA task leverage an attention mechanism of some kind. In our project, we explore the effectiveness of augmenting three different models which use attention with character-level and semantic embeddings, and examine the qualitative effects on their attention behaviors.

Dataset

We conducted our experiments on the SQuAD 2.0 Dataset



Question: Why was Tesla returned to Gospic?
Context paragraph: On 24 March 1879, Tesla was returned to Gospic under police guard for not having a residence permit. On 17 April 1879, Milutin Tesla died at the age of 60 after contracting an unspecified illness (although some sources say that he died of a stroke). During that year, Tesla taught a large class of students in his old school, Higher Real Gymnasium, in Gospic.
Answer: not having a residence permit

Figure 1. Example of SQuAD Question-Context-Answer triple

- Context paragraphs sourced from Wikipedia. Questions and answers crowdsourced using Amazon Mechanical Turk
- Answers are a span from the context paragraph
- Roughly half the questions cannot be answered. Model should return N/A

R-NET

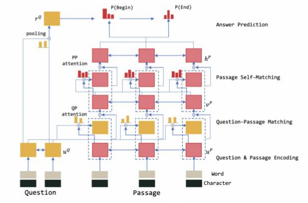


Figure 2. R-NET Architecture

R-NET [1] introduces multiple innovations on attention to augment a core recurrent structure.

- **Gated Attention** A variation of normal attention, gated attention uses a sigmoid gate to mask out parts of both the input passage and attention map, allowing a model to learn to focus its computation.
- **Self-Attention** Effective question answering involves an understanding of question-context correspondence and relationships of multiple pieces of information within the context. Self-attention expresses the relationship between the different parts of the context.

QANET

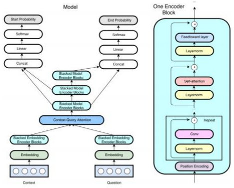


Figure 3. QANet Architecture

QANet [3] innovates on previous work by using a completely attention-based architecture that forgoes recurrent structure for convolutional layers.

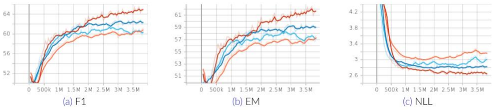
- **Encoder Block** Uses depthwise separable convolutions and multi-headed attention layers to create meaningful representations instead of recurrent layers. It is significantly faster and more memory-efficient than a similarly-configured LSTM or GRU.

Experiments

BiDAF, R-NET, and QANet are three models which use attention at different complexity levels for SQuAD. We experiment with adding three different semantic augmentations.

1. **Character-Level Embeddings**, added to the baseline BiDAF model, are implemented using GRU layers applied over the character embeddings in each word token. Character embeddings can provide richer characterizations of words, especially those which are out of the model's known vocabulary.
2. **Part of Speech**. POS tags were added to all model inputs. Adding Part-of-Speech tags to the model's input may have an effect on its understanding of a word's semantic meaning and its interactions with other tokens by giving it additional expert knowledge, as opposed to hoping it learns these relationships on its own.
3. **Named Entity Recognition**. NER tags were added to the BiDAF model. Named entities include anything that can be denoted with a proper name, such as a person, place, date, etc. We hoped that NER tags would assist the model with identifying key tokens and spans in the context that would be relevant to the question, especially "who", "when", and "where" questions.

Training



(a) F1

(b) EM

(c) NLL

Figure 4. Training scores of BiDAF (orange), BiDAF-POS (cyan), R-NET (blue), and QANet (red)

Results

Model	F1	EM	AvNA
BiDAF Baseline	61.22	57.69	68.06
BiDAF Char Embeddings	62.59	59.17	68.96
BiDAF POS	61.23	58.12	67.90
BiDAF Char Embeddings POS	63.14	59.94	69.8
BiDAF NER	60.83	57.52	67.92
BiDAF POS NER	61.75	58.44	68.21
R-NET	62.66	59.47	69.27
R-NET POS	64.23	60.83	70.71
QANet	65.45	**62.41**	71.55
QANet POS (dev)	**65.62**	**62.41**	**71.79**
QANet POS (test)	62.88	59.73	-

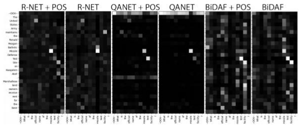


Figure 5. Attention visualizations with and without POS labels

Analysis

- Character embeddings performance increase can be attributed to improved understanding of out-of-vocabulary words.
- POS tags improved performance modestly; QANet virtually unaffected due to its complex question-context representation
  - Likely eventually learned a proxy for POS tags.
- POS tags may have acted as a regularizer for smaller models.

Conclusion

- Character-level embeddings can provide significant benefits to models that initially only contain word-embeddings
- POS tags can act as an effective regularizer for smaller models, but diminishes in return for complex models.

References

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

[2] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi.
Bidirectional attention flow for machine comprehension, 2018.

[3] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le.
Qanet: Combining local convolution with global self-attention for reading comprehension, 2018.