SQuAD 2.0 QA with BiDAF++ & QANet

We produced a question answering system on SQuAD 2.0 by exploring enhancements to improve a baseline BiDAF[1] model. We present results from the addition of character-level embeddings and token features in context encoding, as well as components from the transformer-based QANet[2] architecture.

AUTHORS
Mei Tan | Raymond Zhang
Graduate School of Education
Stanford University

INTRODUCTION

The task of question answering (QA) is an interesting and meaningful area of investigation not only because of the popularity in machine learning research communities but also in the potential of applying such systems across domains into the social sciences. Through this project we hope to gain an understanding of state-of-the-art neural network architectures, effects of different tuning parameters, model evaluation, and input feature decisions. Our ultimate goal is to be able to gain the competency to apply such techniques appropriately to the education domain.

DATA

SQuAD 2.0 dataset consisting of (context, question, answer) triples
Train set of 129,941 examples (43498 unanswerable)
Dev set of 6078 examples (3168 unanswerable)
Test set 5915 examples

Bigram	Count	Percentage
What is	716	8.65
What was	433	5.23
How many	309	3.73
What did	303	3.66
When did	275	3.32
In what	199	2.40
When was	176	2.13
What are	168	2.03
What does	166	2.01
Who was	154	1.86

Unigram	Count	Percentage
What	3811	46.04
Who	752	9.09
How	743	8.98
When	561	6.78
Where	338	4.08
In	328	3.96
Which	237	2.86
The	197	2.38
Why	160	1.93
By	44	0.53

EVALUATION

EM (Exact Match if system output catches ground truth)
F1 (harmonic mean of precision and recall)

EXPERIMENT DETAILS

All models trained for 30 epochs with dropout rate of 0.2 and learning rate of 0.2. Most experiments used batch size 64 and hidden size 100.
Due to equipment constraints, QANet was run with batch size 16 and hidden size of 128.

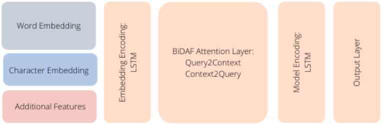
REFERENCES

[1] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. CoRR, abs/1611.01603, 2016.
[2] 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. CoRR, abs/1804.09541, 2018.
[3] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. CoRR, abs/1704.00051, 2017.

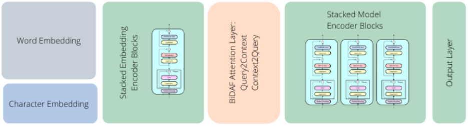
METHODOLOGY

We augment GloVe word vector representation in the embedding layer:

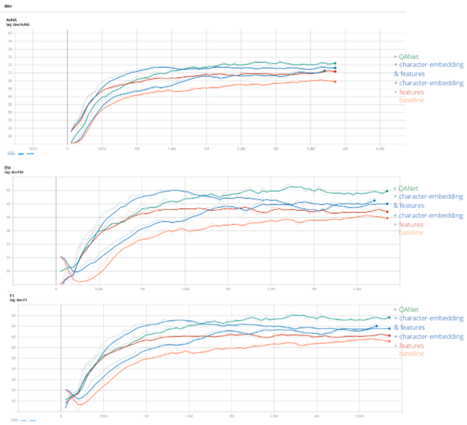
- Character-level embeddings added to context and question words
- Additional token features added to context words (whether a word can be matched to a question word, part-of-speech tag, and entity type) [3]



We also implement the transformer-based QANet architecture with convolution and self attention encoding blocks.


Word Embedding
Character Embedding
Stacked Embedding Encoder Blocks
BiDAF Attention Layer: Query2Context Context2Query
Stacked Model Encoder Blocks
Output Layer

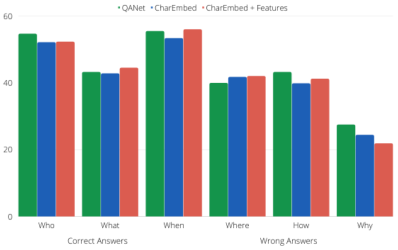
TRAINING PERFORMANCE



ANALYSIS

Dev set performance:

Model	EM	F1
Baseline	58.461	61.696
Baseline + Features	59.57	62.788
Baseline + Char Embed	60.797	64.459
Baseline + Char Embed + Features	62.174	65.275
QANet	62.998	66.547



Ans Wrd Cnt	N	Percentage
0	2090	56.64
1	612	16.59
2	426	11.54
3	244	6.61
4	142	3.85

Ans Wrd Cnt	N	percentage
0	1016	23.00
2	724	16.39
1	713	16.14
3	550	12.45
4	316	7.15

Our models are better at predicting AvNA than more complex answers. However, at times these mistakes are similar to human behavior, as in this example:

Question: Of Poland's inhabitants in 1901, what percentage was Catholic?
Context: Throughout its existence, Warsaw has been a multi-cultural city. According to the 1901 census, out of 711,988 inhabitants 56.2% were Catholics, 35.7% Jews, 5% Greek orthodox Christians and 2.8% Protestants. Eight years later, in 1909, there were 281,754 Jews (36.9%), 18,189 Protestants (2.4%) and 2,818 Mariavites (0.4%). This led to construction of hundreds of places of religious worship in all parts of the town. Most of them were destroyed in the aftermath of the Warsaw Uprising of 1944. After the war, the new communist authorities of Poland discouraged church construction and only a small number were rebuilt.
Answer:N/A
Prediction:56.2%

CONCLUSION

Though the QANet architecture was overall the best performing, including additional input features to represent words more expressively was surprisingly effective. Most importantly through this project we learned several model architectures and gained practice with neural network concepts.