# Question Answering on SQuAD 2.0

Qiwen Wang, Mengyu Li, Boyao Sun
{qwang26, lmy18, boysun}@stanford.edu

## INTRODUCTION

➤ **Machine Comprehension Task**: Comprehend a given passage of text and correctly generates the answer from the context span to the questions.

➤ **SQuAD 2.0 challenge:** Unanswerable questions are added. Models with pre-trained contextual embeddings (PCE) reaches human performance, but requires significant computation power.

**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

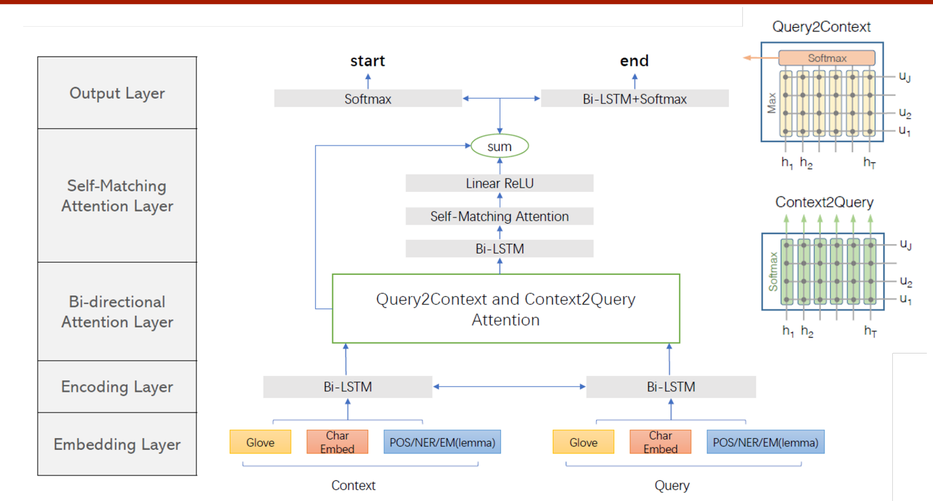
## DATASET AND FEATURES

➤ **Dataset:** SQuAD 2.0

➤ **Dataset size:** training set, dev set, and test set contain 129,941, 6078, and 5915 paragraph and question pairs, over half of the questions start with 'what'.

➤ **Evaluation Metrics:**
F1, EM and AvNA.

## APPROACH




Our model is consisted of 5 layers shown above. We made several modifications to the original BiDAF model to improve its performance, including adding **character-level embeddings** and **word features** in the embedding layer, **applying self-attention** layer, changing the type of RNN layer, optimizer, and learning rate to improve the result.

## Experiment and Result

➤ Achieves **60.626 EM, 63.931 F1** on the test set.

➤ Character embedding and self attention significantly boost the performance

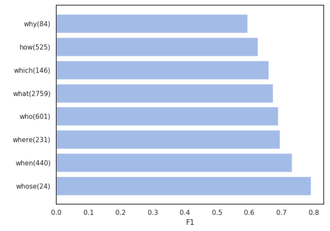
➤ POS/NER is not very effective as features

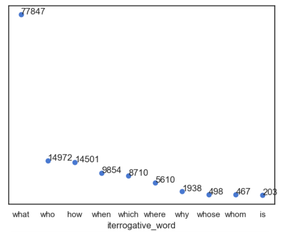
Table 1: Model Performance on SQuAD v2.0 dev set (Non-PCE)

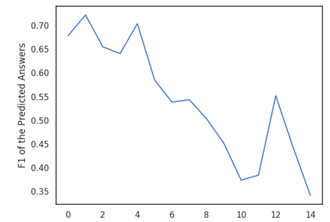
| BiDAF Model | F1 | EM | AvNA |
|---|---|---|---|
| Baseline | 58.64 | 55.57 | 65.18 |
| With char emb, self attention* | 61.77 | 61.54 | 69.25 |
| With char emb, exact match, POS/NER, self attention* | 64.49 | 61.55 | 71.38 |
| With char emb | 64.00 | 60.71 | 70.06 |
| With char emb, exact match | 64.69 | 61.54 | 71.20 |
| With char emb, exact match, self attention | **66.74** | **63.57** | **73.20** |

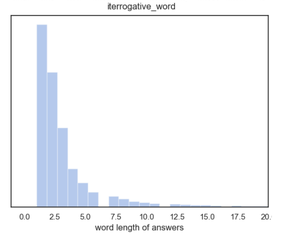
*: Due to the computational limit on Azure NV6, we added a linear layer followed by ReLu activation after the attention layer to reduce the hidden layer dimension.

## Analysis




➤ Higher F1 score for "whose", "when" type of question. NER distinguishes time and the name of a person

➤ "How", "why" type of question needs more complicated logic to answer.

➤ Performs better for answer with short length. The majority number of answers are short.

**Context**: By far the most famous work of Norman art is the Bayeux Tapestry, which is not a tapestry but a work of embroidery. ...
**Question**: What is the oldest work of Norman art?
**Answer**: N/A
**Prediction**: Bayeux Tapestr

➤ Falsely predict the answer whose question resembles to the content.