Problem/Background
As one of the ultimate goals of natural language processing (NLP), machine comprehension can be assessed by answering one or multiple questions with a chunk of text, such as a news article or a short biography. Most benchmark datasets contain questions whose answers are single entities or single tokens, while the Stanford Question Answering Dataset contains questions whose answers can be any sequence of tokens from the passage.

Datasets
- **SQUAD 2.0** developed by Rajpurkar et al. [1]
  - 50,000 unanswerable questions written adversarially
  - Requires the model to determine whether the question is answerable, and answer the question when it is possible
- spacy "en_core_web_sm" WordNet
- The named entity and part of speech recognition are powered by spacy's small English model based on WordNet 3.0
- English Word Frequency dataset
- Published on Kaggle 1/3 million most common English Words

Model Overview

---

Model Details
We propose BIDAF Pro Max for machine reading comprehension tasks based on BIDAF [2]. Besides the basic model,
- we employ new token features with another token feature embed layer, and
- adopt two iterative reattention blocks before attention flow layer.

**Results/Conclusions**

**Experiments**

<table>
<thead>
<tr>
<th>F1</th>
<th>EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIDAF (baseline)</td>
<td>65.60</td>
</tr>
<tr>
<td>BIDAF + Cond. Prob.</td>
<td>61.73</td>
</tr>
<tr>
<td>QANet</td>
<td>55.03</td>
</tr>
<tr>
<td>QANet + Asl. Output Layer</td>
<td>66.13</td>
</tr>
<tr>
<td>BIDAF(C)</td>
<td>65.54</td>
</tr>
<tr>
<td>BIDAF(C) + Token Features</td>
<td>75.95</td>
</tr>
<tr>
<td>BIDAF(C) + Token Features + Iter. Attn.</td>
<td>80.32</td>
</tr>
</tbody>
</table>

*EM/F1 scores improved BIDAF + Character Embedding.

The experiments above show that
- Token features significantly improves the prediction. Compared with BIDAF(C), token features raise F1 score and EM score by more than 10 points.
- The iterative attention further improves the model.

- The absolute game changer is EM token feature. Single feature of EM can boost the F1 score from 65.54 to 77.25.
- Although the ablation study shows No TF token feature configuration outperforms that with four, four token features and iterative attentions, the model works better than without TF.

Reference