Neural Question Answering at BioASQ 5B

Georg Wiese, Dirk Weissenborn, Mariana Neves
Motivation

- **Neural question answering (QA) systems** are end-to-end trainable machine learning models which achieve top performance in domains with **large training datasets**

- We apply an **extractive neural QA system** (FastQA [1]) to BioASQ 5B Phase B (list & factoid questions)

- **Extractive QA**: Answer is given as start and end pointers in the context (snippets)
Network Architecture

Original FastQA [1]  Our Architecture

Start Probabilities $p_{start}$  
End Probabilities $p_{end}$

Start Probabilities $p_{start}$  
End Probabilities $p_{end}$

Core FastQA

Start Scores $y_{start}$  
End Scores $y_{end}$  

Start Scores $y_{start}$  
End Scores $y_{end}$

GloVe Embeddings  Character Embeddings

GloVe Embeddings  Character Embeddings

Biomedical Embeddings  Question Type Features

Context Embeddings  Question Embeddings

Context Embeddings  Question Embeddings
Network Architecture

Input Layer
- GloVe & character embeddings (like the original FastQA)
- Biomedical embeddings [3]
- Question type features
Network Architecture

Output Layer

- Change start probability activation from softmax to **sigmoid**
- **Multiple starts** can be selected for list questions
- For each selected start, select the corresponding end pointer via softmax

Original FastQA [1]  Our Architecture
Network Architecture

Original FastQA [1]  Our Architecture
Training Procedure

- **Problem**: Neural QA typically requires $\sim 10^5$ questions to train

- Datasets of such scale exist in the open domain, e.g. SQuAD [2] with $\sim 10^5$ factoid questions on Wikipedia articles

- We train in two steps:
  1. Pre-training on a large ($\sim 10^5$ questions) open-domain dataset (SQuAD)
  2. Fine-tuning on BioASQ ($\sim 10^3$ questions)
Systems

- We trained **five models** using 5-fold cross validation on all available training data.

- We submitted **two systems**:
  - **Single**: Best single model according to its respective development set
  - **Ensemble**: Ensemble of all five models (averaging scores before sigmoid/softmax activation)
Results

Factoid Results:
- Our system won 3/5 batches
- Averaged over the five batches, our system (ensemble) was 1.5 percentage points above the best competitor

<table>
<thead>
<tr>
<th>Batch</th>
<th>Best Competitor</th>
<th>Single</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40.0% (LabZhu-FDU)</td>
<td>52.0%</td>
<td>57.1%</td>
</tr>
<tr>
<td>2</td>
<td>48.4% (LabZhu-FDU)</td>
<td>38.3%</td>
<td>42.6%</td>
</tr>
<tr>
<td>3</td>
<td>38.5% (LabZhu-FDU)</td>
<td>43.1%</td>
<td>42.1%</td>
</tr>
<tr>
<td>4</td>
<td>32.1% (LabZhu-FDU)</td>
<td>29.7%</td>
<td>36.1%</td>
</tr>
<tr>
<td>5</td>
<td>42.4% (LabZhu-FDU)</td>
<td>39.2%</td>
<td>35.1%</td>
</tr>
<tr>
<td>Average</td>
<td>40.3%</td>
<td>39.7%</td>
<td>41.8%</td>
</tr>
</tbody>
</table>
Results

List Results:

- Our system won 2/5 batches
- On average, the best competitor performed 3.4 percentage points better than our ensemble model

<table>
<thead>
<tr>
<th>Batch</th>
<th>Best Competitor</th>
<th>Single</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.3% (BioASQ_Baseline)</td>
<td>33.6%</td>
<td>33.5%</td>
</tr>
<tr>
<td>2</td>
<td>50.0% (LabZhu-FDU)</td>
<td>29.0%</td>
<td>26.2%</td>
</tr>
<tr>
<td>3</td>
<td>39.0% (LabZhu-FDU)</td>
<td>41.5%</td>
<td>49.5%</td>
</tr>
<tr>
<td>4</td>
<td>37.5% (LabZhu-FDU)</td>
<td>24.2%</td>
<td>29.3%</td>
</tr>
<tr>
<td>5</td>
<td>41.0% (LabZhu-FDU)</td>
<td>36.1%</td>
<td>39.1%</td>
</tr>
</tbody>
</table>

Average | 39.2%     | 33.4%  | 35.8%    |
Discussion

Strengths: Competitive performance, despite:

- Less feature engineering than traditional QA systems
- A less domain-dependent architecture, because we don’t rely on domain-specific structured resources

Limitations:

- Extractive QA cannot generate answer which are not explicitly mentioned in the snippets
  - No yes/no & summary questions
References

[1] Weissenborn et al.: “Making Neural QA as Simple as Possible but not Simpler”


Thank You. Questions?

Related CONLL paper:
“Neural Domain Adaptation for Biomedical Question Answering”

Contact:
georg.wiese@student.hpi.de