Neural Question Answering at BioASQ 5B

Georg Wiese, Dirk Weissenborn, Mariana Neves

German Research Center for Artificial Intelligence

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

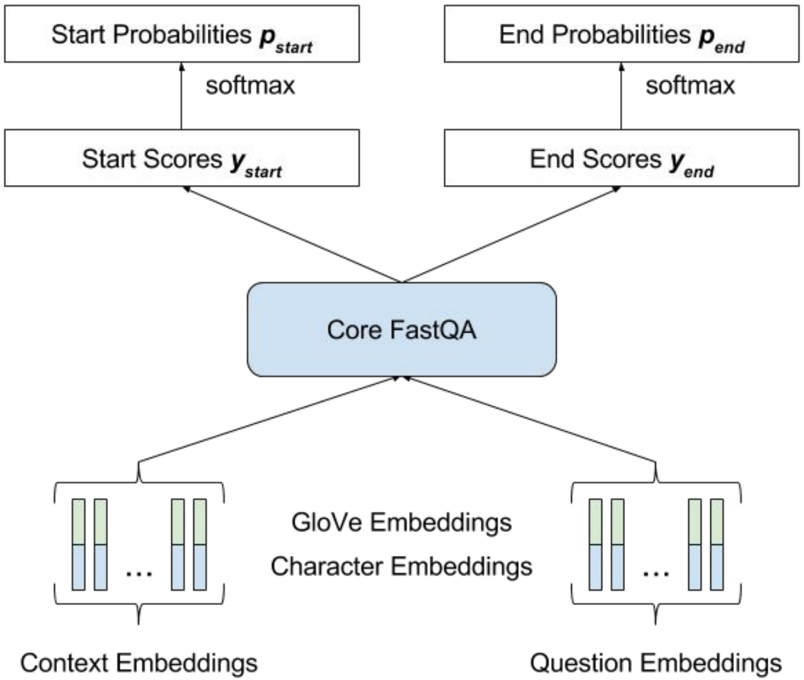
Motivation

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

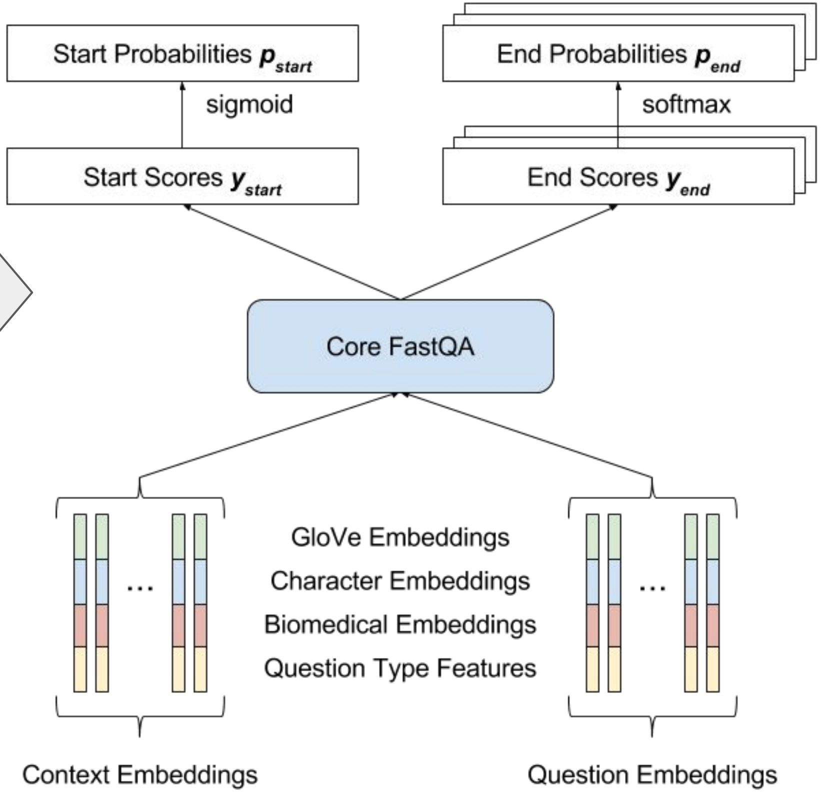
- We apply an **extractive neural QA** system (FastQA [1]) to BioASQ 5B Phase B (list & factoid questions)

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

Network Architecture



Original FastQA [1]

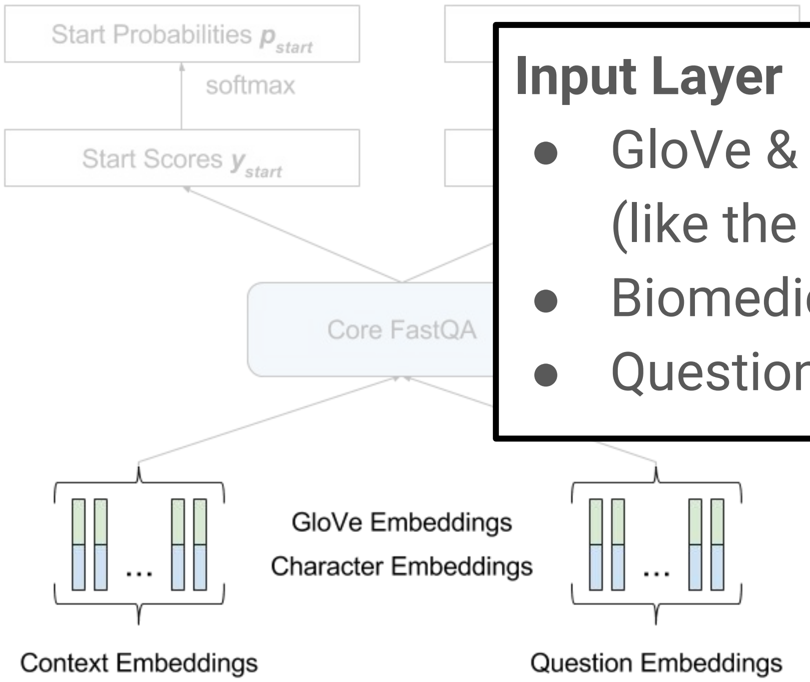


Our Architecture

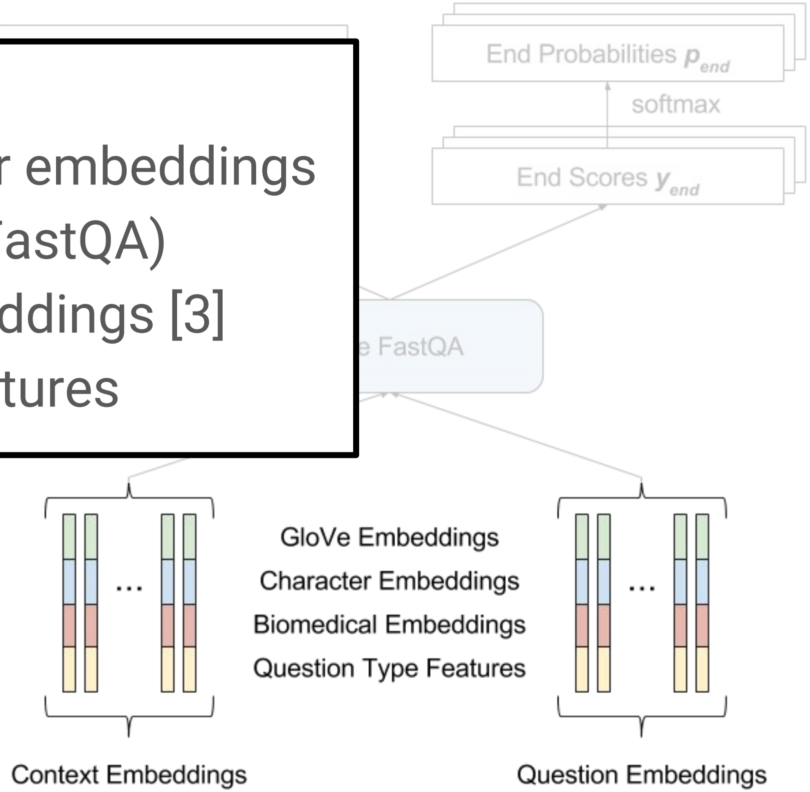
Network Architecture

Input Layer

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

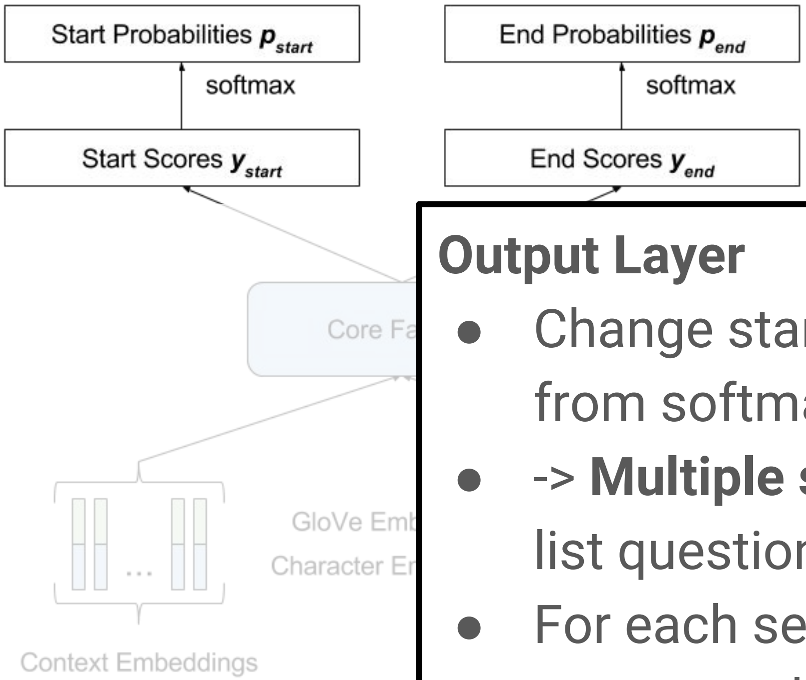


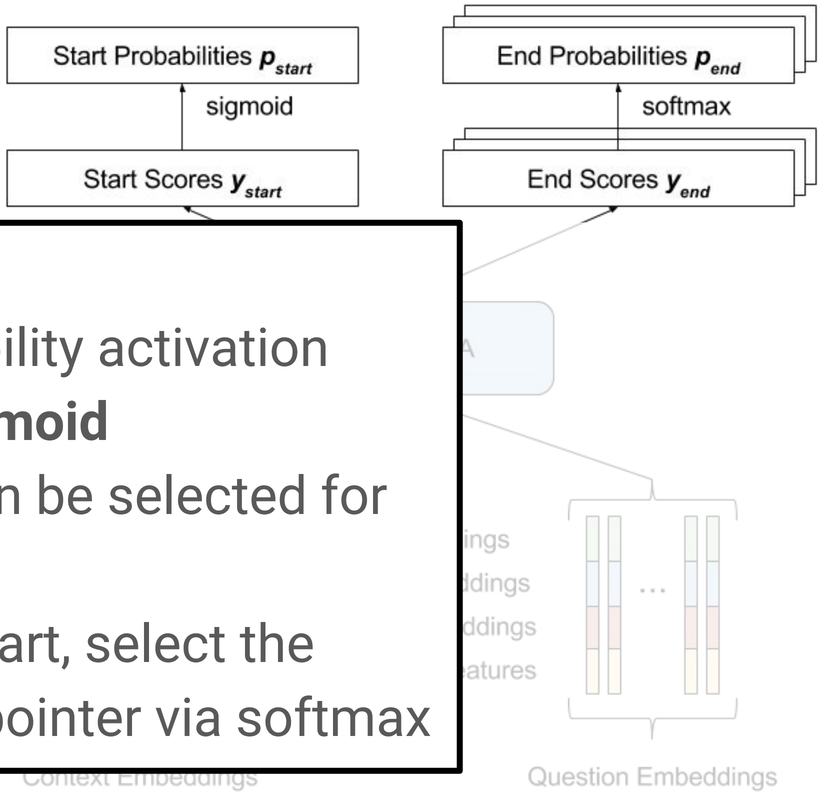
Original FastQA [1]



Our Architecture

Network Architecture





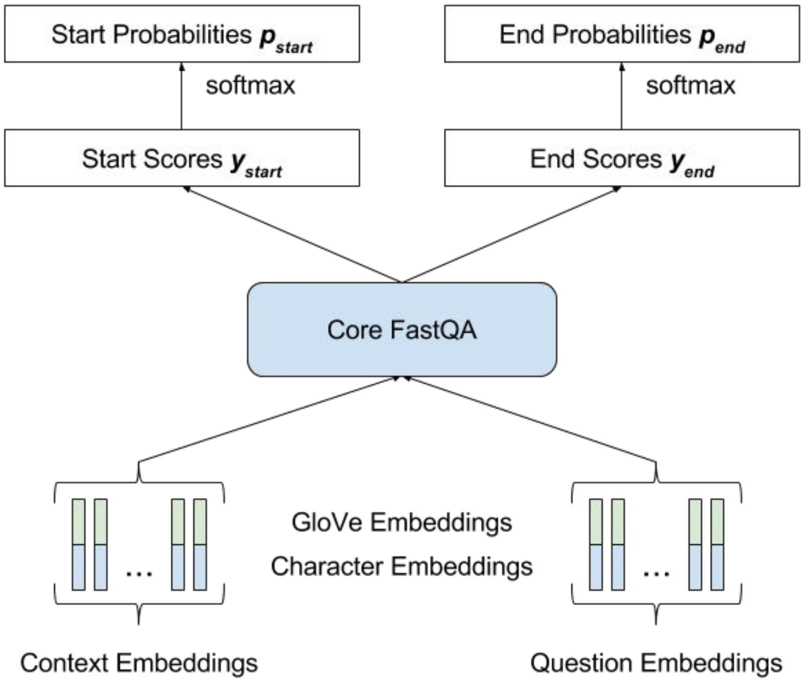
Output Layer

- 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

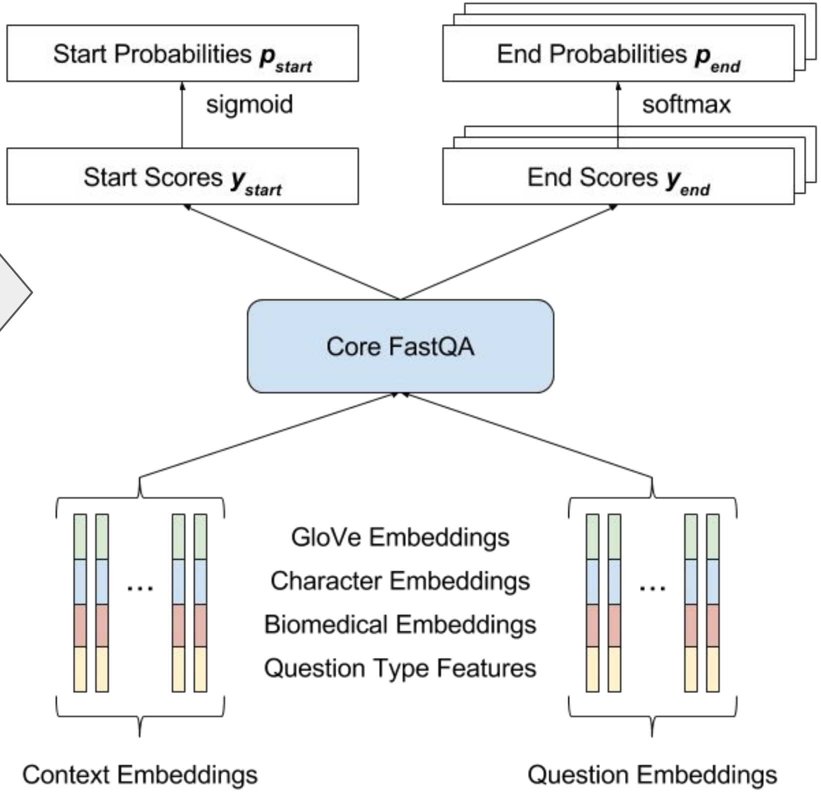
Original FastQA [1]

Our Architecture

Network Architecture



Original FastQA [1]



Our Architecture

Training Procedure

- **Problem**: Neural QA typically requires ~10^5 questions to train
- Datasets of such scale exist in the open domain, e.g. SQuAD [2] with ~10^5 factoid questions on Wikipedia articles
- We train in two steps:
  1. Pre-training on a large (~10^5 questions) open-domain dataset (SQuAD)
  2. Fine-tuning on BioASQ (~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

Batch	Best Competitor	Single	Ensemble
1	40.0% (LabZhu-FDU)	52.0%	**57.1%**
2	**48.4%** (LabZhu-FDU)	38.3%	42.6%
3	38.5% (LabZhu-FDU)	**43.1%**	42.1%
4	32.1% (LabZhu-FDU)	29.7%	**36.1%**
5	**42.4%** (LabZhu-FDU)	39.2%	35.1%
Average	40.3%	39.7%	**41.8%**

Results

List Results:

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

Batch	Best Competitor	Single	Ensemble
1	31.3% (BioASQ_Baseline)	**33.6%**	33.5%
2	**50.0%** (LabZhu-FDU)	29.0%	26.2%
3	39.0% (LabZhu-FDU)	41.5%	**49.5%**
4	**37.5%** (LabZhu-FDU)	24.2%	29.3%
5	**41.0%** (LabZhu-FDU)	36.1%	39.1%
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”

[2] Rajpurkar et al.: “SQuAD: 100,000+ Questions for Machine Comprehension of Text”

[3] Pavlopoulos et al.: “Continuous Space Word Vectors Obtained by Applying Word2Vec to Abstracts of Biomedical Articles”

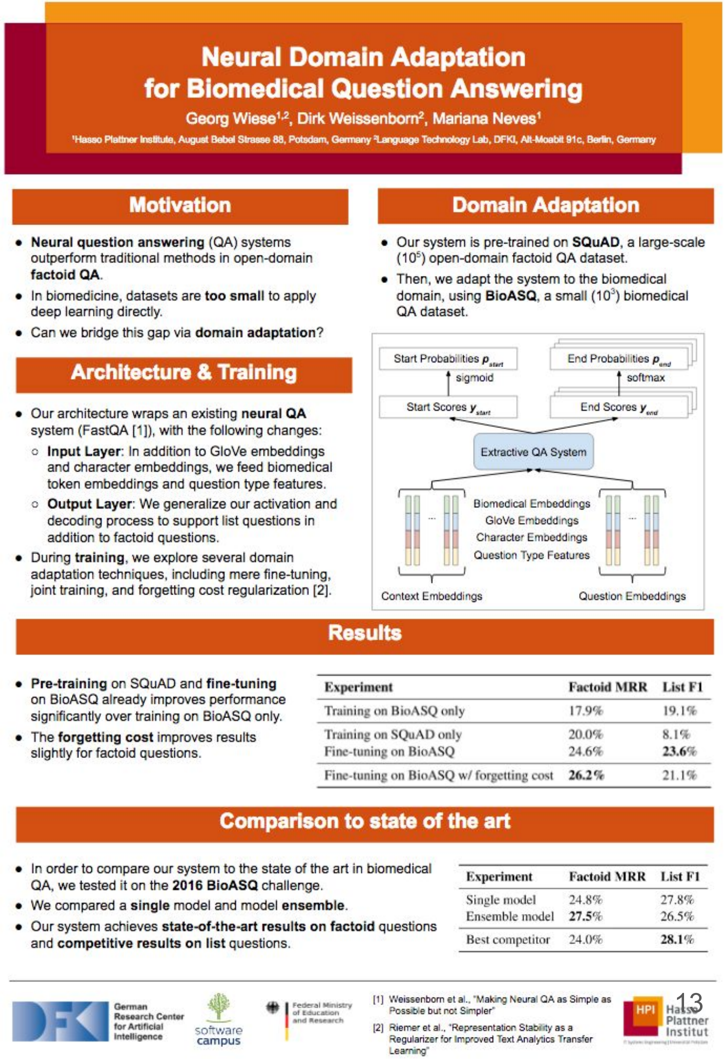
Thank You. Questions?

Related CONLL paper:

“Neural Domain Adaptation for Biomedical Question Answering”

Contact:

georg.wiese@student.hpi.de



Neural Domain Adaptation for Biomedical Question Answering

Georg Wiese[1,2], Dirk Weissenborn[2], Mariana Neves[1]

[1]Hasso Plattner Institute, August Bebel Strasse 88, Potsdam, Germany [2]Language Technology Lab, DFKI, Alt-Moabit 91c, Berlin, Germany

Motivation

- **Neural question answering** (QA) systems outperform traditional methods in open-domain **factoid QA**.
- In biomedicine, datasets are **too small** to apply deep learning directly.
- Can we bridge this gap via **domain adaptation**?

Domain Adaptation

- Our system is pre-trained on **SQuAD**, a large-scale (10^5) open-domain factoid QA dataset.
- Then, we adapt the system to the biomedical domain, using **BioASQ**, a small (10^3) biomedical QA dataset.

Architecture & Training

- Our architecture wraps an existing **neural QA** system (FastQA [1]), with the following changes:
  - **Input Layer**: In addition to GloVe embeddings and character embeddings, we feed biomedical token embeddings and question type features.
  - **Output Layer**: We generalize our activation and decoding process to support list questions in addition to factoid questions.
- During **training**, we explore several domain adaptation techniques, including mere fine-tuning, joint training, and forgetting cost regularization [2].

Results

- **Pre-training** on SQuAD and **fine-tuning** on BioASQ already improves performance significantly over training on BioASQ only.
- The **forgetting cost** improves results slightly for factoid questions.

Experiment	Factoid MRR	List F1
Training on BioASQ only	17.9%	19.1%
Training on SQuAD only	20.0%	8.1%
Fine-tuning on BioASQ	24.6%	**23.6%**
Fine-tuning on BioASQ w/ forgetting cost	**26.2%**	21.1%

Comparison to state of the art

- In order to compare our system to the state of the art in biomedical QA, we tested it on the **2016 BioASQ** challenge.
- We compared a **single** model and model **ensemble**.
- Our system achieves **state-of-the-art results on factoid** questions and **competitive results on list** questions.

Experiment	Factoid MRR	List F1
Single model	24.8%	27.8%
Ensemble model	**27.5%**	26.5%
Best competitor	24.0%	**28.1%**

German Research Center for Artificial Intelligence
software campus
Federal Ministry of Education and Research

[1] Weissenborn et al., "Making Neural QA as Simple as Possible but not Simpler"
[2] Riemer et al., "Representation Stability as a Regularizer for Improved Text Analytics Transfer Learning"

HPI Hasso Plattner Institut