Neural Domain Adaptation for Biomedical Question Answering

Georg Wiese\textsuperscript{1,2}, Dirk Weissenborn\textsuperscript{2}, Mariana Neves\textsuperscript{1}

\textsuperscript{1}Hasso Plattner Institute, August Bebel Strasse 88, Potsdam, Germany
\textsuperscript{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?

Architecture & Training

- Our architecture wraps an existing neural QA system (FastQA \cite{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 \cite{2}.

Domain Adaptation

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

\begin{itemize}
  \item Pre-training on SQuAD and \textbf{fine-tuning} on BioASQ already improves performance significantly over training on BioASQ only.
  \item The \textbf{forgetting cost} improves results slightly for factoid questions.
\end{itemize}

Results

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|}
\hline
Experiment & Factoid MRR & List F1 \\
\hline
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\% \\
\hline
\end{tabular}
\end{table}

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 \textbf{single} model and model \textbf{ensemble}.
- Our system achieves \textbf{state-of-the-art results on factoid} questions and \textbf{competitive results on list} questions.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|}
\hline
Experiment & Factoid MRR & List F1 \\
\hline
Single model & 24.8\% & 27.8\% \\
Ensemble model & 27.5\% & 26.5\% \\
Best competitor & 24.0\% & 28.1\% \\
\hline
\end{tabular}
\end{table}

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