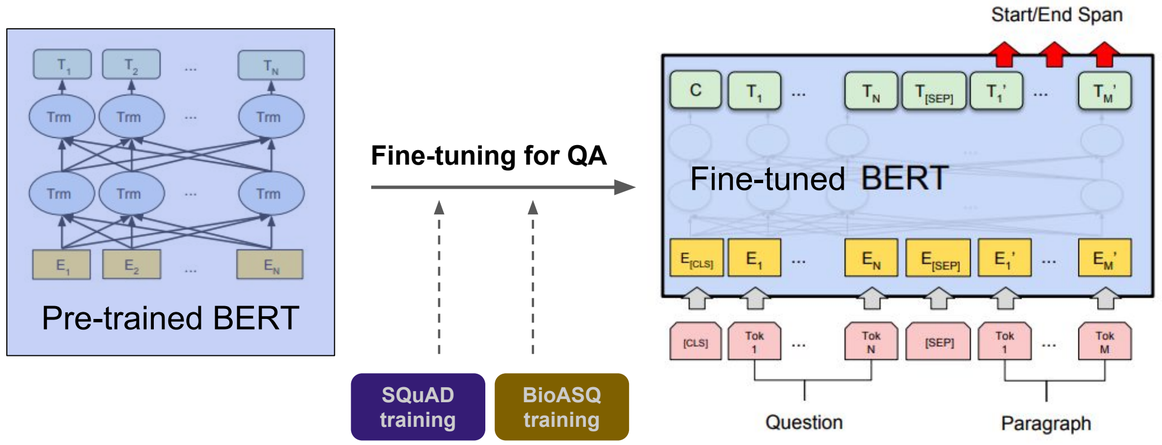
SQuAD to BioASQ: analysis of general to specific

Karen Ouyang | kjouyang@stanford.edu
Biomedical Informatics

MOTIVATION

As biomedical information in the form of publications and electronic health records (EHR) increases at an increasingly fast pace, there is clear utility in having systems that can automatically handle information extraction, summarization, and question answering tasks. While there have been significant strides in improving language tasks for general language, addressing domain-specific contexts still remains challenging. In this project, I apply and fine-tune models to the SQuAD dataset and further modify and adapt for biomedical domain-specific question answering. I evaluated and compared performance on the SQuAD dataset and BioASQ, a biomedical literature QA dataset, with the goal of analyzing and developing approaches to leverage unsupervised language models for domain-specific applications. Upon generating various fine-tuned models, the best performance for general language SQuAD QA achieved an F1 score of 76.717, EM score of 73.379, and for biomedical-specific BioASQ QA achieved an F1 score of 70.348 and EM score of 49.902.

MODEL ARCHITECTURE



DATA

SQuAD2.0 The Stanford Question Answering Dataset

- Train set: 129,941 examples
- Dev set: 6,078 examples

CONTEXT: " . . . Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the **Bald Eagle Protection Act** of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised."
QUESTION: "What was the name of the 1937 treaty?"
ANSWER: "Bald Eagle Protection Act"

BioASQ

- Train set: 24,559 examples
- Dev set: 6,140 examples

CONTEXT: "... Among the miRNAs that showed a consistent regulation tendency through all specimens and showed more than a 2-fold difference in serum, 5 miRNAs (miR-132, miR-26a, let-7b, miR-145, and miR-143) were determined as the 5 most markedly down-regulated miRNAs in the serum from ovarian cancer patients with respect to those of controls. Four miRNAs (miR-132, miR-26a, let-7b, and miR-145) out of 5 selected miRNAs were significantly underexpressed in the serum of ovarian cancer patients in qRT-PCR.\n\nCONCLUSIONS: **Serum miR-132, miR-26a, let-7b, and miR-145 could be considered as potential candidates as novel biomarkers in serous ovarian cancer.** Also, serum miRNAs is a promising and useful tool for discriminating between controls and patients with serous ovarian cancer."
QUESTION: "Which miRNAs could be used as potential biomarkers for epithelial ovarian cancer?"
ANSWER: "Serum miR-132, miR-26a, let-7b, and miR-145 could be considered as potential candidates as novel biomarkers in serous ovarian cancer"

RESULTS & ANALYSIS

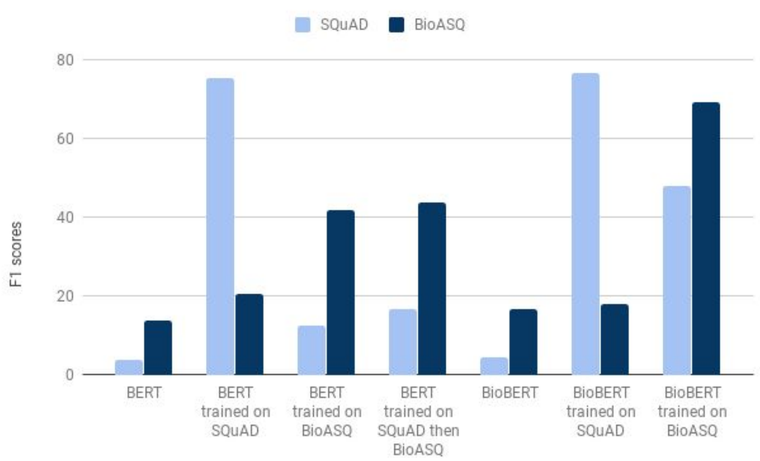
Example of predicted BioASQ QA sample

- **Question:** Is there an association between Klinefelter syndrome and breast cancer?
- **Context:** There appear to be no substantial data to confirm the assumption that breast cancer in men with Klinefelter's syndrome is as common as breast cancer in the normal female population. The number of reported cases of breast cancer in Klinefelter's males is only 27, a number too small for any meaningful statistical analysis. There is evidence, however, to suggest that Klinefelter's males have an increased risk of breast cancer that approaches three percent. Physicians should therefore be aware of potential breast pathology in XXY males and incorporate a careful breast examination and specific education into the routine health maintenance of men with Klinefelter's syndrome.
- **Answer:** There is evidence, however, to suggest that Klinefelter's males have an increased risk of breast cancer that approaches three percent.
- **Prediction:** Klinefelter's males have an increased risk of breast cancer that approaches three percent.

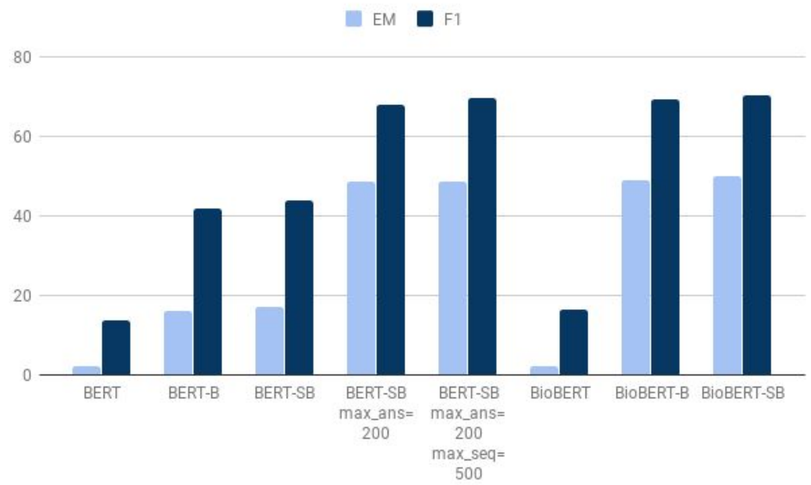
Best performing models for SQuAD and BioASQ QA

Model name	Description	EM	F1
BERT-S	SQuAD dev set	71.997	75.349
BioBERT-S	SQuAD dev set	73.379	76.717
BERT-SB, max-ans=200, max-seq=500	BioASQ dev set	48.615	69.503
BioBERT-SB, max-ans=200, max-seq=500	BioASQ dev set,	49.902	70.348

Comparison of SQuAD and BioASQ performance



Comparison of Models Predicting BioASQ QA



CONCLUSION

These results demonstrate that leveraging an unsupervised language model, BERT, for domain-specific QA with substantially less supervised training data achieves results that are nearing comparable to general language QA.