Biomedical Question Answering with BioBERT and SDNet

CS 224N: Natural Language Processing with Deep Learning course project

Lu Yang, Erin Brown, Sophia Lu
Stanford University

Motivation

- Over **800,000** new citations added to MEDLINE in 2017 alone PubMed PubMed Central
- Not feasible for medical professionals to keep up with recent developments
- Aim to develop a novel, contextual QA system specifically for biomedical-related text mining

Related Work

- **SDNet** : a contextualized attention-based deep network designed for conversational QA
- **BioBERT** : a domain specific language representation model based on BERT and pre-trained on large-scale biomedical corpora

Approach

- Extractive factoid question answering
- Adapt SDNet for non-conversational QA
- Integrate BioBERT with SDNet
- Compare against BioBERT alone
- Data provided by the **BioASQ** challenge

BioASQ Example

Context: Disruption of ALX1 causes extreme microphthalmia and severe facial clefting: expanding the spectrum of autosomal-recessive ALX-related **frontonasal dysplasia.**

Question: Which disease has been associated to a disruptive ALX1 protein?

References

[1] Tsatsaronis et al., "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," BMC Bioinformatics, 2015. [2] Zhu, Zeng, and Huang, "SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering," arXiv, 2018. [3] Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," arXiv,, 2019.

SDNet-BioBERT

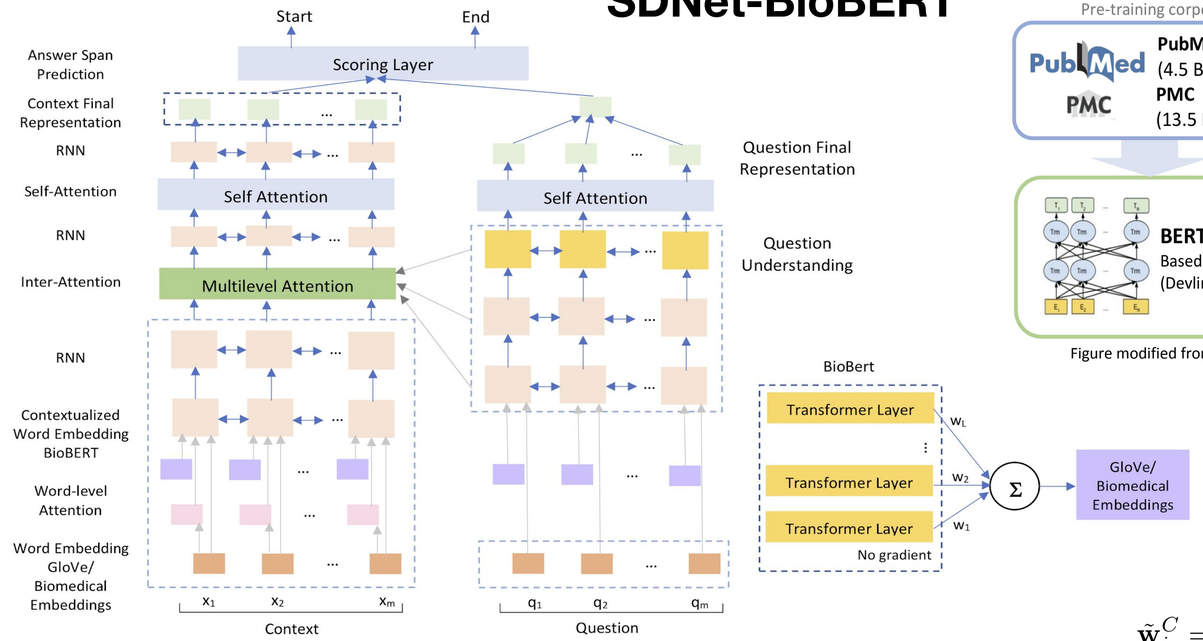


Figure modified from [2]

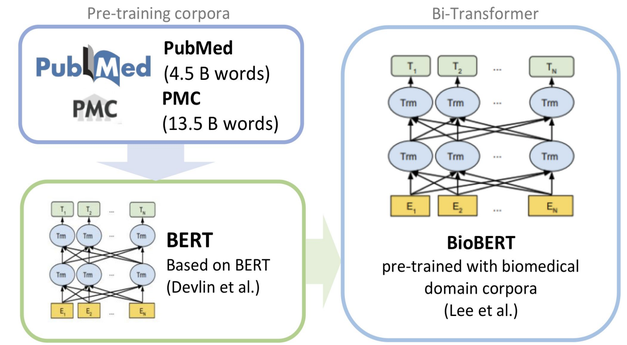


Figure modified from [3]

Word Vectors

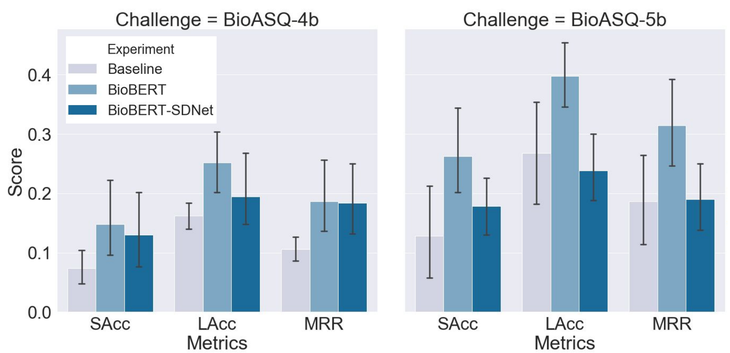
$\hat{\mathbf{h}}_i$ Attended vectors (question to context)

$\mathbf{f}_{w_i}$ Feature vector including POS & NER embedding, indicator, normalized term frequency

$$\tilde{\mathbf{w}}_i^Q = \left[\text{GloVe}_{w_i^Q}; \text{BioBERT}_{w_i^Q}\right]$$

$$\tilde{\mathbf{w}}_i^C = \left[\text{GloVe}_{w_i^C}; \text{BioBERT}_{w_i^C}; \hat{\mathbf{h}}_i^C; \mathbf{f}_{w_i^C}\right]$$

Results



Question 1: "Where are Paneth cells located?"
Golden: "in the intestinal crypt base columnar cells"
BioBERT-SDNet: "Intestinal stem cells"
BioBERT: "The intestinal epithelium is a classic example of a rapidly self-renewing tissue fueled by dedicated resident stem cells."

Question 2: "In which yeast chromosome does the rDNA cluster reside?"
Golden: "Chromosome XII"
BioBERT-SDNet: "Chromosome XII"
BioBERT: "."

51.4% of BioBERT predictions and **57.1%** of BioBERT-SDNet predictions achieve partial match with golden answer

	Stories	Literature	School	News	Wikipedia	Overall
DrQA+PGNet	64.2	63.7	67.1	68.3	71.4	65.1
BioBERT-SDNet	**72.0**	**66.5**	**67.7**	**71.3**	**75.4**	**70.5**
BERT-SDNet	75.4	73.9	77.1	80.3	83.1	78.0
Human	90.2	88.4	89.9	88.6	89.9	88.8

F1 scores of our BioBERT-SDNet predictions on CoQA

Conclusion

- Both BioBERT and BioBERT-SDNet outperformed the published baseline
- Quantitatively, BioBERT performed better than the combined model
- BioBERT-SDNet offers some qualitative improvements in answer generation that mark it as a promising area for future study
- Further work including pre-training on a larger dataset such as SQuAD is needed