Contextual embedding and model weighting by fusing domain knowledge on Biomedical Question Answering

Yuxuan Lu, Jingya Yan, Zhixuan Qi, Zhongzheng Ge, Yongping Du

Beijing University of Technology

August 10, 2022

Table of Contents

About me

About me

Yuxuan Lu (卢雨轩)
Senior student at Beijing University of Technology
Intern at MSRA & LinkedIn
luyuxuanleo@gmail.com

Introduction

Question Answering

- Classic task in Natural Language Processing
- Goal:
  - *automatically* extract useful information from a massive number of literatures efficiently
- Applications:
  - Search Engines
  - Automated Customer Service
  - Conversational AI: Siri, Google Assistant, etc.
- QA in Biomedical Domain:
  - Help search for papers and answers.

Challenges of domain-specific QA – 1

Lack of large-scale supervised datasets for most tasks

- Labeling supervised biomedical dataset is hard.
- BioASQ8: 3,243 questions
- SQuAD 1.1: 100,000+ questions

Challenges of domain-specific QA – 2

More knowledge is required to be learned by the model

- Open-domain QA: **Common Sense**
  - Apples are red.
- Domain-Specific QA: Common Sense, **Domain Terms**, **Domain Knowledge**
  - Hyponatremia is a low sodium concentration in the blood.

Challenges of domain-specific QA – 3

Severe Out-Of-Vocabulary problem.

- Open-domain:
  - new words
  - relatively small number
- Biomedical domain:
  - combination of existing word roots
    - hypo-, -natr-, -emia
  - usually large number

Method

Key Point

Combining Bio-BERT and AoA Reader

- AoA Reader: Perform well in **open-domain** QA
- BioBERT: **Unsupervised pre-trained** on large **biomedical** corpus

MLP-based weighting strategy

- learn the preference and biases of two models

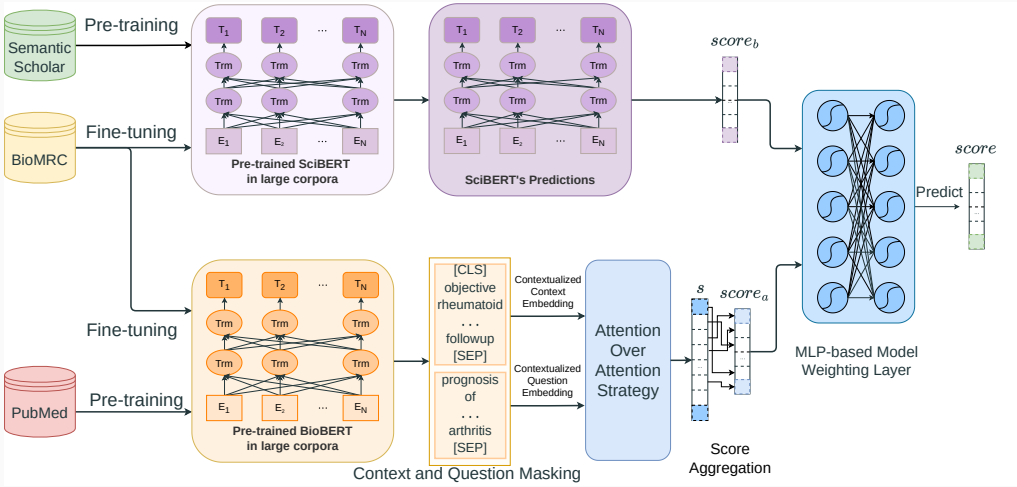


Figure 1: Model structure based on Pre-training and Weighting Strategy

Our answer to those challenges

Lack of large-scale supervised datasets

- We use the BIOMRC dataset
  - 100,000 **Cloze-style** questions
  - relatively larger, but still not enough

Learning of **Domain Terms** and **Domain Knowledge**

- Bio-BERT pre-trained language model
  - **Unsupervised pre-trained** on large-scale biomedical corpus

Our answer to those challenges

Out-Of-Vocabulary words

- WordPiece
  - Thanks to BERT
  - Segment words into ”subwords”
    - we know the subword ”hypo”, ”natr”, ”emia”
    - ”hyponatremia” is represented as ”hypo”, ”##natr”, ”##emia”
  - Matches how these words are made
    - Many domain terms are made by combining existing word roots.

Formal Task Description

- Cloze-style QA questions.
- A question $\langle \mathcal{C}, \mathcal{Q}, \mathcal{A} \rangle$:
  - Context: $\mathcal{C} = \{w_1, w_2, \ldots, w_n\}$
  - Question: $\mathcal{Q} = \{q_1, q_2, \ldots, [MASK], \ldots, q_m\}$
  - Answer Candidates: $\mathcal{A} = \{a_1, a_2, \ldots, a_o\}$
- learn a function F to predict the optimal answer of the question $\mathcal{Q}$ based on the context $\mathcal{C}$:

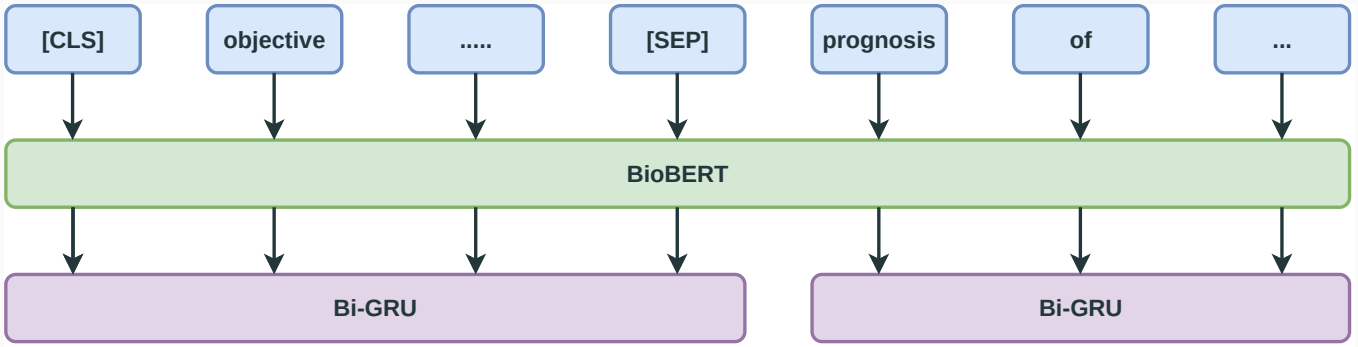
$$\forall a \in \mathcal{A}, P(a|\mathcal{C}, \mathcal{Q}) = \begin{cases} 1 & a \text{ is the correct answer} \\ 0 & a \text{ is not the correct answer} \end{cases} \tag{1}$$

$$F(\mathcal{C}, \mathcal{Q}, \mathcal{A}) = \underset{a \in \mathcal{A}}{argmax}\, P(a|\mathcal{C}, \mathcal{Q}) \tag{2}$$

Training of SciBERT

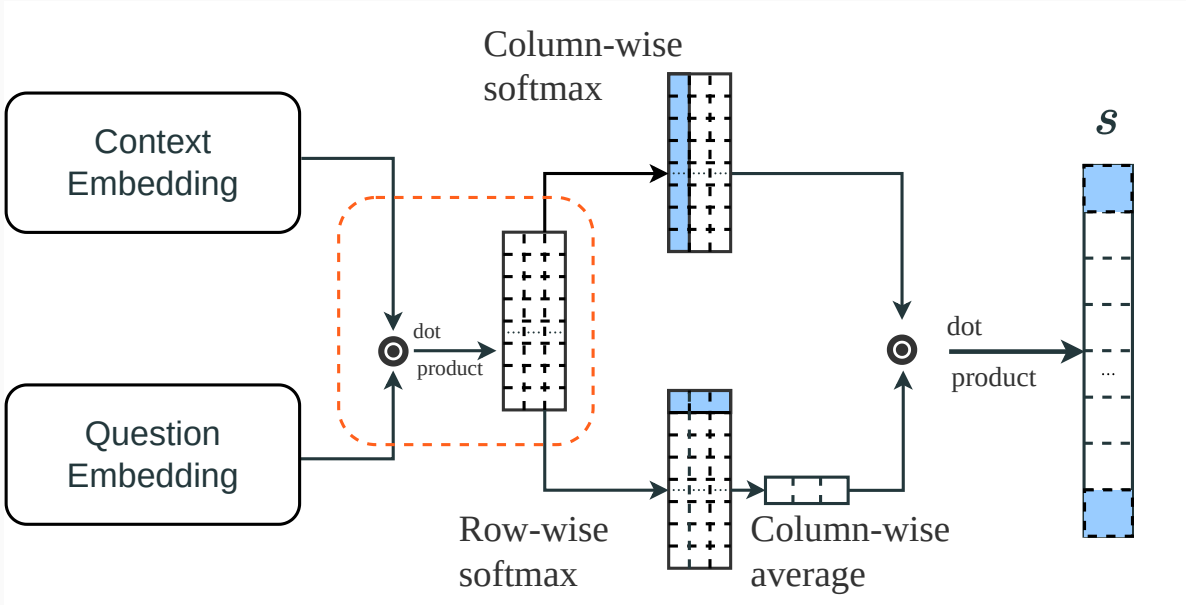
- Training data
  - Pre-train on Semantic Scholar corpus
    - 18% computer science domain
    - 82% biomedical domain
  - Fine-tune on BIOMRC
- Answer Extraction
  - Use same strategy Pappas et al. used in their BIOMRC dataset's baseline
  - Concatenate each context question with query
  - Fed into SciBERT
  - Obtain score through MLP

Contextual word embedding



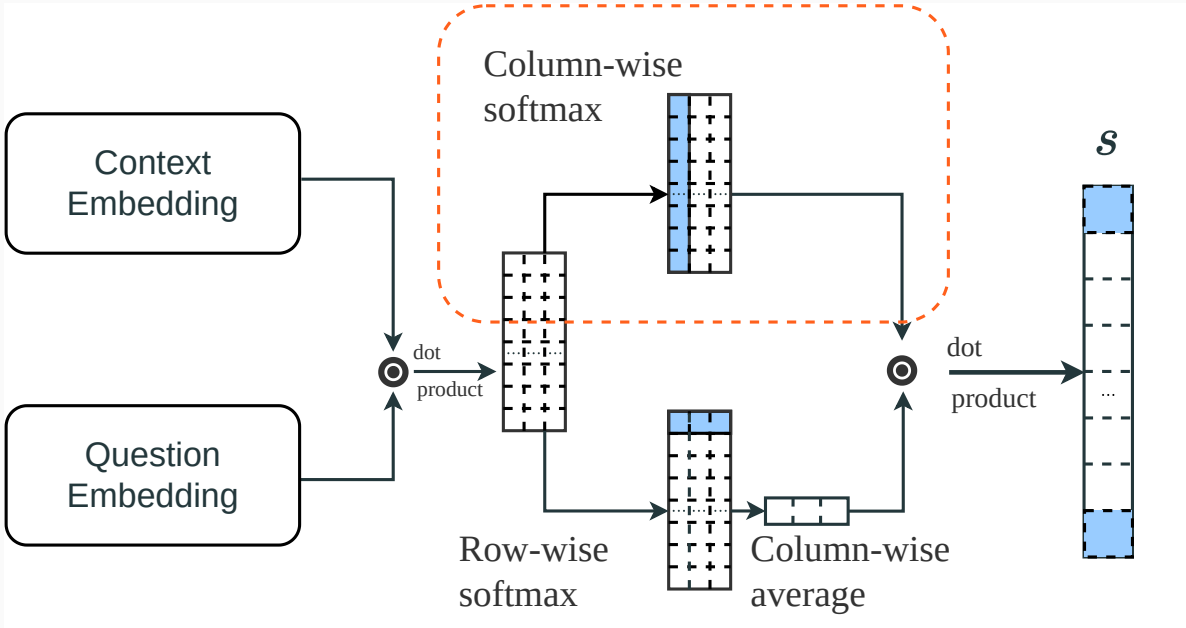
- Use BioBERT to generate contextual word embedding
  - obtain domain-oriented knowledge and terms
- WordPiece is used for tokenization
  - Matches word formation methods
  - Solve Out-Of-Vocabulary issue
- Apply Bi-GRU on context and question **separately** to further obtain contextual representations

Pair-wise Matching Score and Attentions over Attention Mechanism



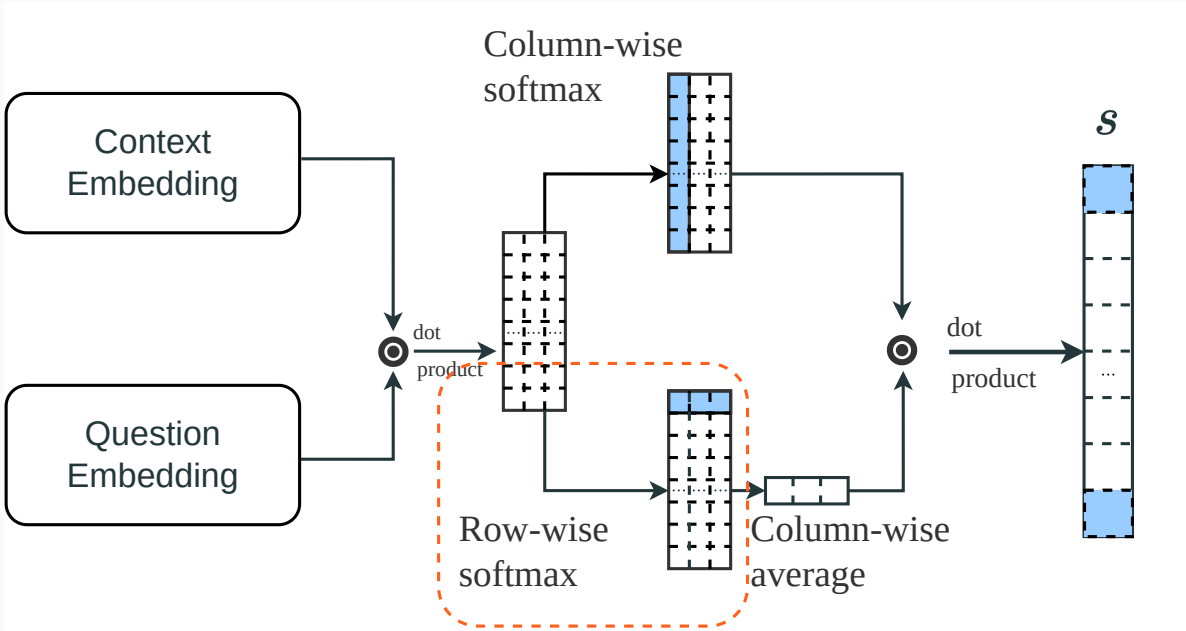
- Calculate a pair-wise matching matrix
  - indicating relevance between tokens in context and question

Pair-wise Matching Score and Attentions over Attention Mechanism



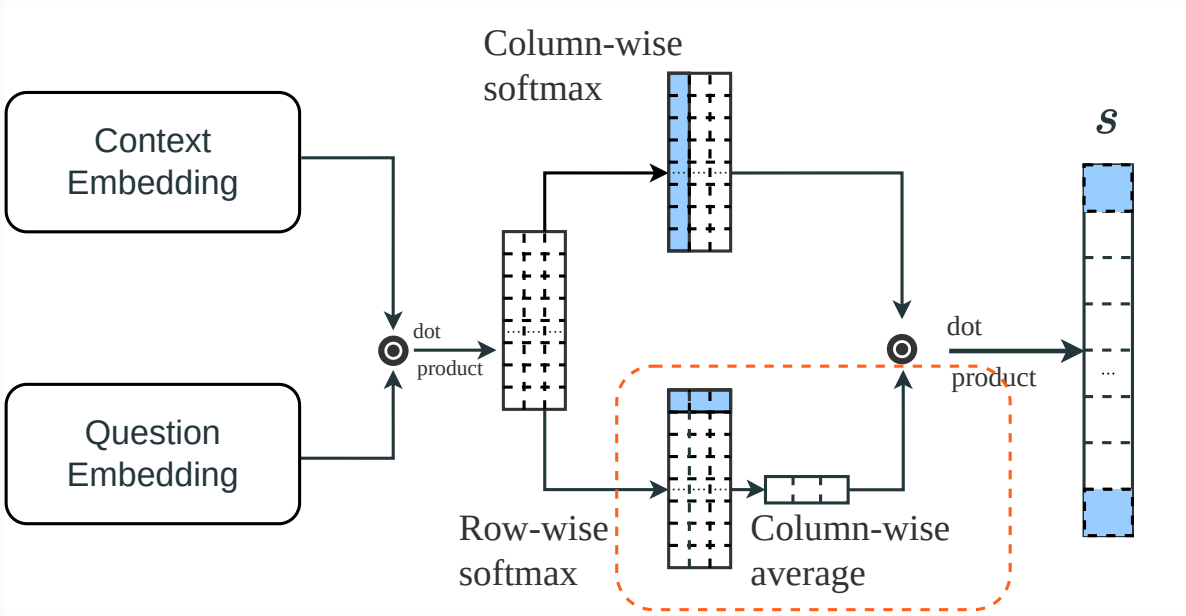
- Column-wise softmax
  - Get context-level attention regarding each token in the question

Pair-wise Matching Score and Attentions over Attention Mechanism



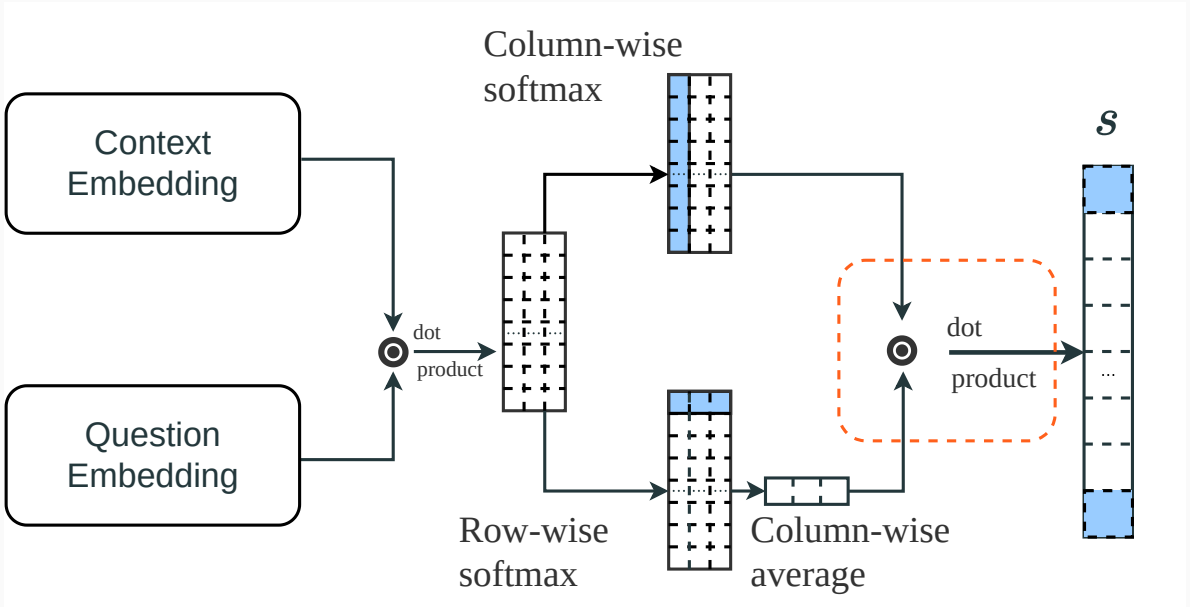
- Row-wise softmax
  - Obtain the ”importance” of each in token the question regarding each token in the context

Pair-wise Matching Score and Attentions over Attention Mechanism



- Column-wise average
  - Obtain the importance of each token in the question regarding whole context

Pair-wise Matching Score and Attentions over Attention Mechanism



- Attention-over-Attention mechanism
  - Merge two attentions
  - Attended context-level attention
  - Obtain the score of each token in the context

Final Predictions

- AoA Reader use *sum attention* mechanism, which isn't appliable
- Two-level of aggregation method is needed
  - 1st level: an answer candidate may occur multiple times in the context
  - 2nd level: in each occurrence, the candidate may be segmented into multiple tokens by WordPiece
    - or the candidate is composed of multiple words
- Different aggregation method is experimented in each level
  - Can either be **maximum** or **sum**

Model Weighting Strategy

- Different models have preferences and biases on data
- Different models perform differently against data with different features
- A simple MLP is used to combine the result of two candidate models
  - **automatically** learn the difference between models
  - take full advantage of both models
  - achieve better performance

Experiment

Dataset

- The BIOMRC dataset is used
- Papers from PUBTATOR is used to form biomedical cloze-style questions
- Contexts are the abstract section of papers
- Answer candidates are biomedical entities extracted from the abstract
- Questions are the title of papers
    - Randomly replace a biomedical entity with a placeholder

Dataset

Context	Because of reports of anaplastic transformation following irradiation, this study examines the incidence of anaplastic transformation and local control of these lesions. This review of seven @entity1 who had @entity189 of the @entity135 that was treated with irradiation shows local control in 71% of cases. There were no cases of anaplastic transformation. This report adds to the literature two cases of "de-differentiation" to less differentiated @entity957 ; one such case occurred after surgery alone. The literature is reviewed. Overall, anaplastic transformation is reported in 7% of @entity1 who had irradiation. De-differentiation occurs after surgery as well. The rate of local control with irradiation is less than 50%; with surgery it is 85%. It is concluded that surgery should be used if the procedure has acceptable morbidity. Otherwise, irradiation can be used. Failures can be salvaged surgically. "Anaplastic transformation" should not affect treatment approach.
Candidate Entities	@entity1: ['patients'] @entity135: ['head and neck'] @entity957: ['squamous carcinomas'] @entity189: ['verrucous carcinoma']
Question	Radiotherapy in the treatment of XXXX of the @entity135 .
Answer	@entity189: ['verrucous carcinoma']

Figure 2: A example of the BIOMRC dataset.

Performance of the Contextual Embedding Strategy

Table 1: The result of different aggregation functions, compared to the state-of-the-art model and human experts

Method	Occurrence Aggregation	Token Aggregation	Train Time	BIOMRC LITE		BIOMRC TINY
				Dev Acc	Test Acc	Test Acc
AS-READER	-	-	16.56hr	62.29	62.38	66.67
AoA-READER	-	-	60.90hr	70.00	69.87	70.00
SCIBERT-MAX-READER	-	-	83.22hr	80.06	79.97	90.00
Human Experts	-	-	-	-	-	85.00
AoA-READER with BioBERT Embedding	max	max	1.50hr	78.54	78.11	90.00
	max	sum	0.88hr	83.40	83.36	93.33
	sum	max	3.60hr	80.98	81.20	90.00
	sum	sum	1.76hr	**87.22**	**86.74**	**93.33**

Performance of the Weighting Model

- The SCIBERT-MAX-READER proposed by Pappas et al. is implemented
- The result of SCIBERT-MAX-READER and the AoA Reader is used to train MLP-based weighting layer

Method	BIOMRC LITE		BIOMRC TINY
	Dev Acc	Test Acc	Test Acc
AoA-READER with BioBERT Embedding	87.22	86.74	93.33
SCIBERT-MAX-READER	79.74	80.21	86.67
MLP-based Weighting Model (ours)	**88.76**	**88.00**	**96.66**
The union of two single models (ideal result)	93.07	92.26	96.66

Performance of the Weighting Model

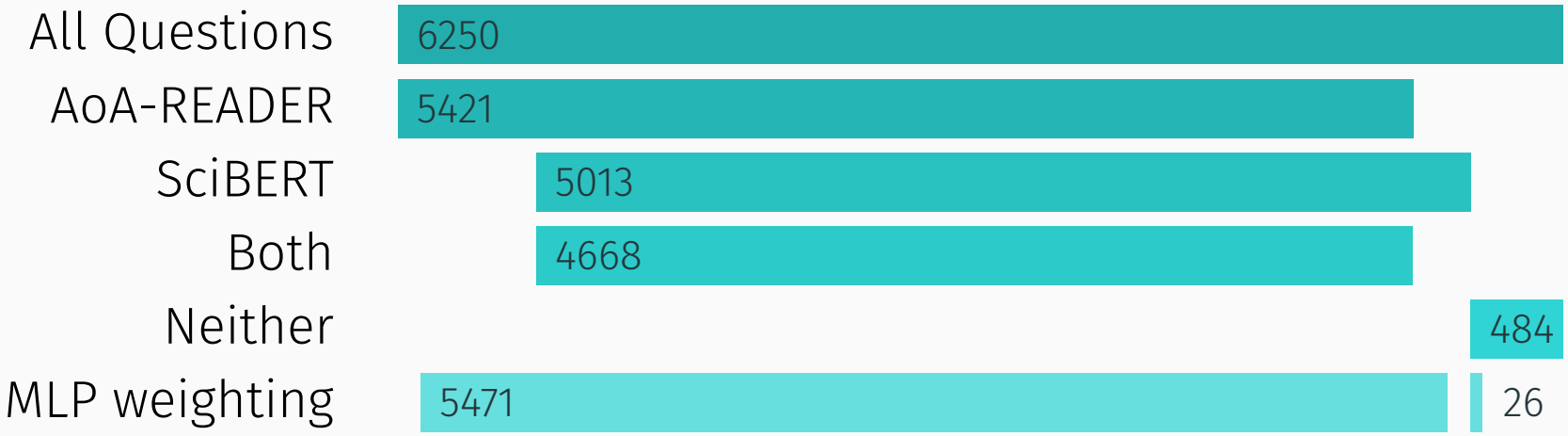


Figure 3: The number of question answered correctly by different models on the BIOMRC LITE dataset.

Conclusion

Conclusion

- Contextual embedding
  - Combine biomedical language model and open-domain QA model
  - mine semantic and contextual information in biomedical QA
- MLP-based model weighting strategy
  - **automatically** learn and utilize the preference and biases of two models
- Outperforms state-of-the-art systems by a large margin
- Achieve higher accuracy than human experts

Thank You!

Any Questions?

Go to https://yuxuan.lu/bcb2022 to get this slide