BioBERT: a pre-trained biomedical language representation model for biomedical text mining

Zixuan Pan (Patrick)

Zitong Li (Cassandra)

UNIVERSITY OF TORONTO

Agenda

- **Motivation & Related Work**
- Model
- Evaluation
- Advantage & Limitation

UNIVERSITY OF
TORONTO

Motivation

- Rapid boost of biomedical information since 2000(eg. PubMed)

- Increasing demand for accurate tools to extract the information from massive biomedical literatures

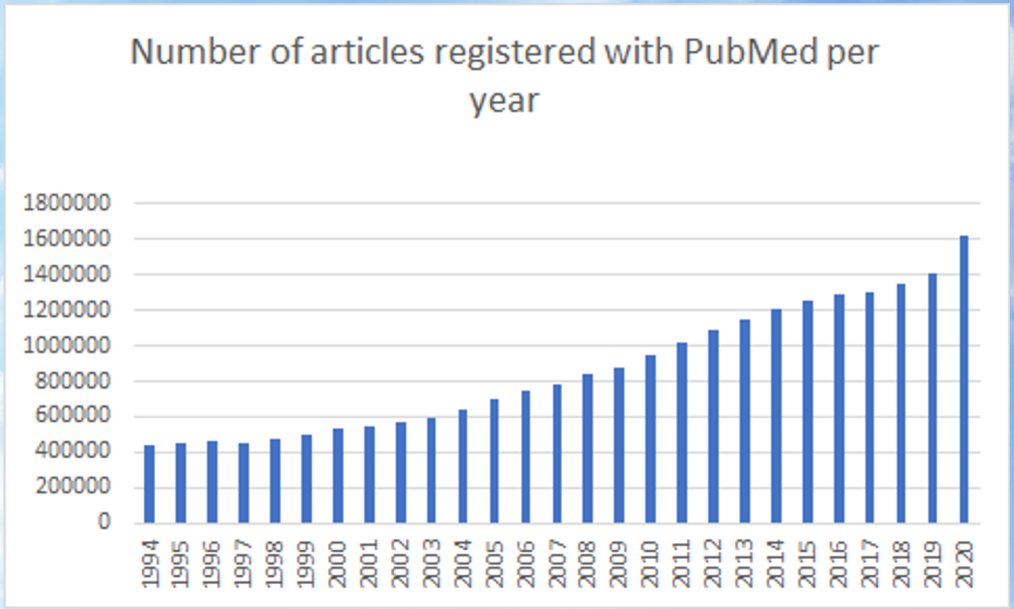


Figure Source: PubMed.gov

UNIVERSITY OF TORONTO

Objective

- Named Entity Recognition (NER)

  Locate and classify the named entity into the pre-determined categories

- Relation Extraction (RE)

  Extract semantic relationship from two or more entities

- Question Answering (QA)

  Extract topics from question provided and generate corresponding answers

Related Work

- Deep learning based models boost up the development of advanced biomedical text mining models
  - Implementing Long Short-Term Memory and Conditional Random Field in Named Entity Recognition [1]
  - Using Recurrent Neural Network architectures to extract chemical-gene relationship from sentences in natural language. [2]
- Limitation:
  - Slow to train, especially when input is a long sequence of words
  - Sequential flow doesn't fully utilize current GPU which are designed for parallel computing

UNIVERSITY OF TORONTO

[1]Giorgi,J.M. and Bader,G.D. (2018) Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics, 34, 4087.
[2]Lim,S. and Kang,J. (2018) Chemical–gene relation extraction using recursive neural network. Database, 2018

Agenda

- Motivation & Related Work
- **Model - BERT**
- Evaluation
- Advantage & Limitation

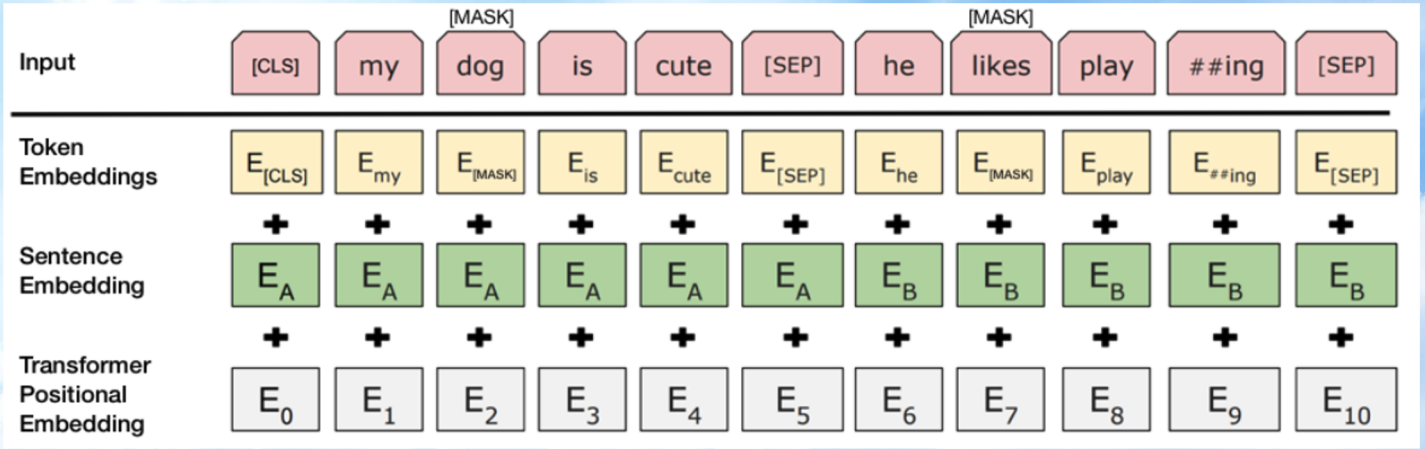
Model: BERT (Bidirectional **E**ncoder **R**epresentation from **T**ransformers**)**

- Multi-layer bidirectional transformer encoder [1]
- Two Steps in the framework: pre-training and fine tuning
  - Pre-training: Train on unlabeled data over different pre-training task
  - Fine-tuning: Train on labeled data from downstream work.
- Highlights:
  - Bidirectional pre-training for language representation
  - Unified framework in both Pre-training and Fine-Tuning
  - Fine-Tuning phase can be used for various natural language processing (NLP) tasks

UNIVERSITY OF TORONTO

[1]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

BERT: Input Representation (How does model understand the words ?)

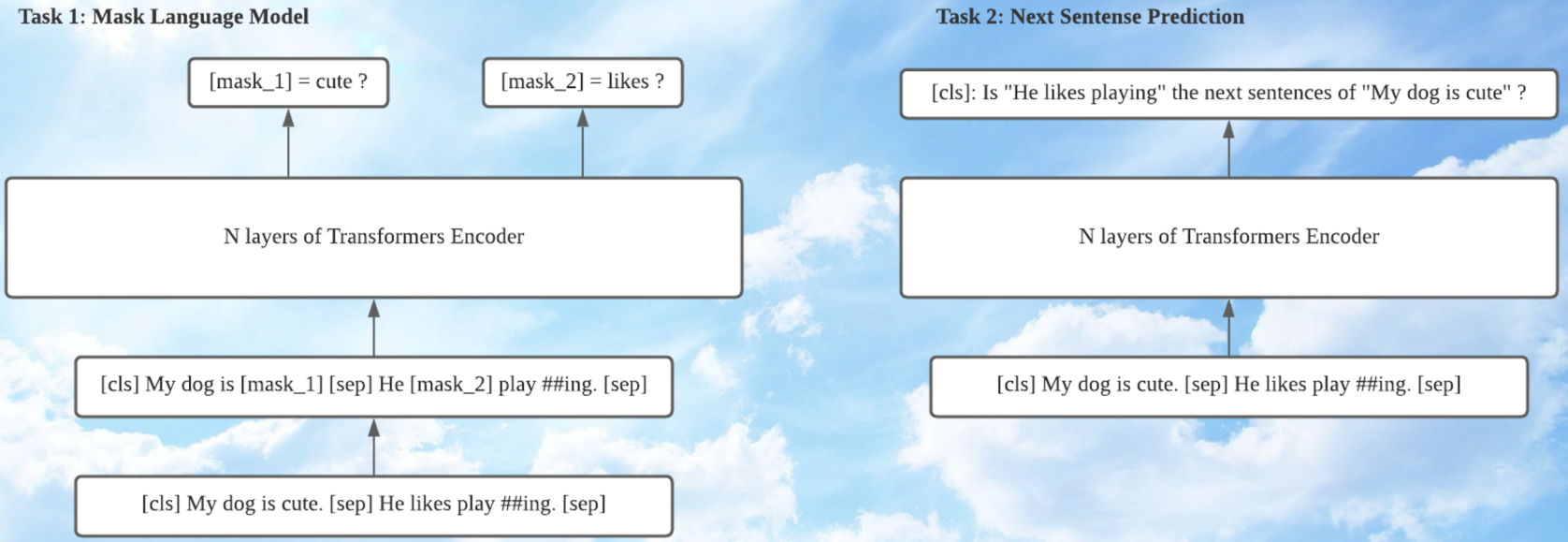


- Token Embedding: Using real-value vectors to encode the meaning of the words
- Sentence Embedding: record the sentence information of each token
- Transformer Positional Embedding: record position information of each token

UNIVERSITY OF TORONTO

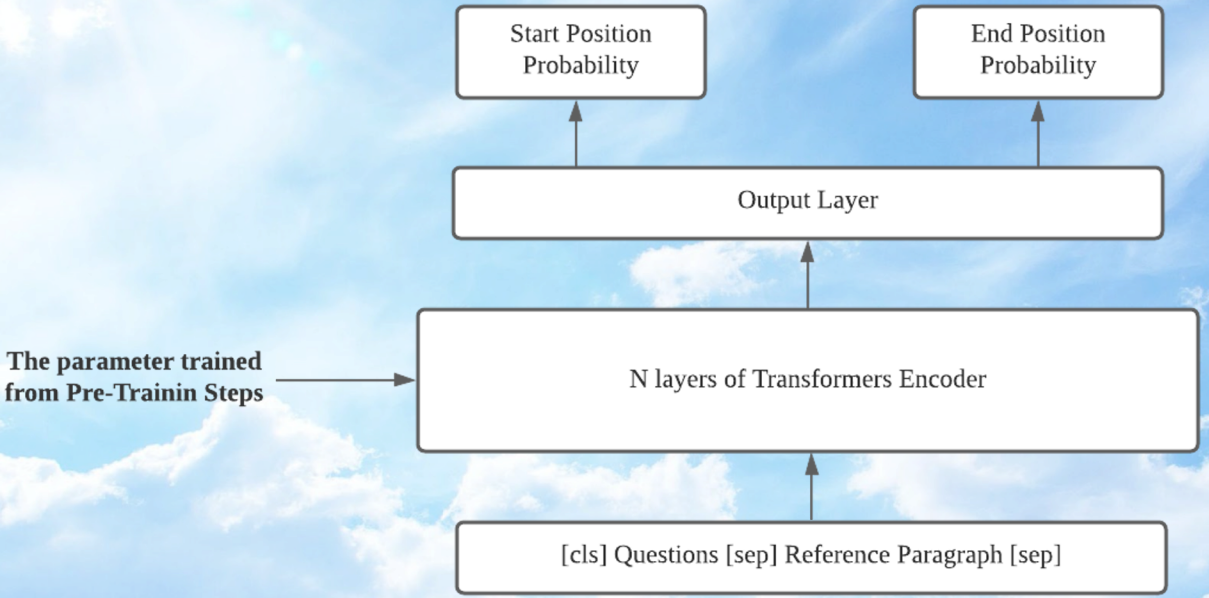
Figure Source: Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

BERT: Pre-Training



- Task 1: Mask LM
  - Mask random token in the input sequence.
  - Final hidden vector of the masked token are fed in to the softmax over the vocabulary
- Task 2: Next Sentence Prediction (NSP)

BERT: Fine-Tuning



- Initialized based on pre-training phase
- Input varies based on the specific task, (e.g. Question and Answering)

UNIVERSITY OF TORONTO

Figure Source: Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Agenda

- Motivation & Related Work
- **Model - BioBERT**
- Evaluation
- Advantage & Limitation

UNIVERSITY OF TORONTO

BioBERT: Pre-Training Dataset

- Other than the datasets used to pre-train BERT, BioBert adds biomedical domain corpora during pre-training
  - PubMed abstracts
  - PMC full-text articles

Table 1. List of text corpora used for BioBERT

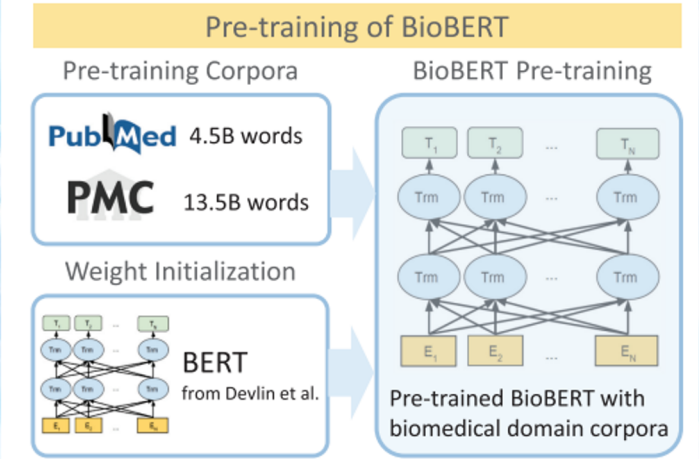
Corpus	Number of words	Domain
English Wikipedia	2.5B	General
BooksCorpus	0.8B	General
PubMed Abstracts	4.5B	Biomedical
PMC Full-text articles	13.5B	Biomedical

UNIVERSITY OF TORONTO

Figure Source: Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published. https://doi.org/10.1093/bioinformatics/btz682

BioBERT: Pre-training

- Initialize as BERT with same weights
- Pre-trained on biomedical domain corpora
- Use WordPiece Tokenization for out-of-vocabulary issue and with case vocabulary for slightly better performance in downstream tasks
  - e.g. Immunoglobulin -> I ##mm ##uno ##g ##lo ##bul ##in



UNIVERSITY OF TORONTO

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published. https://doi.org/10.1093/bioinformatics/btz682

BioBERT: Pre-training Experiment

- BioBERT v1.0 (+ PubMed / + PMC) is the version of BioBERT trained for 470K steps
- After initial release of BioBERT v1.0, pre-trained BioBERT on PubMed for 1M steps, and refer to this version as BioBERT v1.1 (+ PubMed)

Table 2. Pre-training BioBERT on different combinations of the following text corpora: English Wikipedia (Wiki), BooksCorpus (Books), PubMed abstracts (PubMed) and PMC full-text articles (PMC)

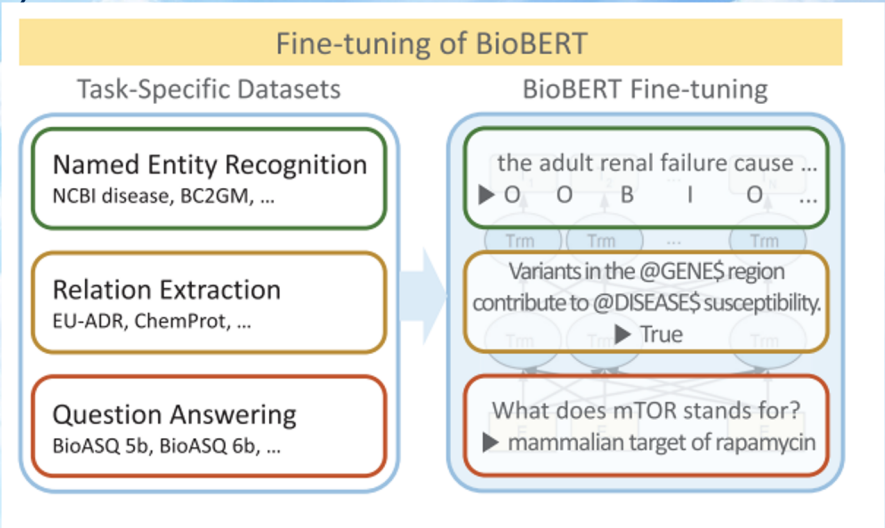
Model	Corpus combination
BERT (Devlin *et al.*, 2019)	Wiki + Books
BioBERT (+PubMed)	Wiki + Books + PubMed
BioBERT (+PMC)	Wiki + Books + PMC
BioBERT (+PubMed + PMC)	Wiki + Books + PubMed + PMC

UNIVERSITY OF TORONTO

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published. https://doi.org/10.1093/bioinformatics/btz682

BioBERT: Fine-Tuning

- Named Entity Recognition (NER)
- Relation Extraction (RE)
- Question Answering (QA)



UNIVERSITY OF TORONTO

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published. https://doi.org/10.1093/bioinformatics/btz682

Fine-Tuning: Named Entity Recognition (NER)

- BioBERT directly learns WordPiece embeddings during pre-training and fi tuning

Table 3. Statistics of the biomedical named entity recognition datasets

Dataset	Entity type	Number of annotations
NCBI Disease (Doğan *et al.*, 2014)	Disease	6881
2010 i2b2/VA (Uzuner et al., 2011)	Disease	19 665
BC5CDR (Li *et al.*, 2016)	Disease	12 694
BC5CDR (Li *et al.*, 2016)	Drug/Chem.	15 411
BC4CHEMD (Krallinger *et al.*, 2015)	Drug/Chem.	79 842
BC2GM (Smith *et al.*, 2008)	Gene/Protein	20 703
JNLPBA (Kim *et al.*, 2004)	Gene/Protein	35 460
LINNAEUS (Gerner *et al.*, 2010)	Species	4077
Species-800 (Pafilis *et al.*, 2013)	Species	3708

Note: The number of annotations from Habibi *et al.* (2017) and Zhu *et al.* (2018) is provided.

UNIVERSITY OF TORONTO

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published. https://doi.org/10.1093/bioinformatics/btz682

Fine-Tuning: Relation Extraction (RE)

- Utilized the sentence classifier of the original version of BERT, which uses a [CLS] token for the classification of relations
- Anonymized target named entities in a sentence using pre-defined tags such as @GENE$ or @DISEASE$
  - Serine at position 986 of @GENE$ may be an independent genetic predictor of angiographic @DISEASE$

Table 4. Statistics of the biomedical relation extraction datasets

Dataset	Entity type	Number of relations
GAD (Bravo et al., 2015)	Gene–disease	5330
EU-ADR (Van Mulligen *et al.*, 2012)	Gene–disease	355
CHEMPROT (Krallinger *et al.*, 2017)	Protein–chemical	10 031

Note: For the CHEMPROT dataset, the number of relations in the training, validation and test sets was summed.

UNIVERSITY OF TORONTO

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published. https://doi.org/10.1093/bioinformatics/btz682

Fine-Tuning: Question Answering (QA)

- Used the same BERT architecture used for SQuAD
- BioASQ datasets are used because their format is similar to that of SQuAD
- Token level probabilities for the start/end location of answer phrases are computed using a single output layer

Table 5. Statistics of biomedical question answering datasets

Dataset	Number of train	Number of test
BioASQ 4b-factoid (Tsatsaronis *et al.*, 2015)	327	161
BioASQ 5b-factoid (Tsatsaronis *et al.*, 2015)	486	150
BioASQ 6b-factoid (Tsatsaronis *et al.*, 2015)	618	161

UNIVERSITY OF TORONTO

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published. https://doi.org/10.1093/bioinformatics/btz682

Agenda

- Motivation & Related Work
- Model
- **Evaluation**
- Advantage & Limitation

UNIVERSITY OF TORONTO

BioBERT: Evaluation - NER

- Matrics:
  - Precision = TP/TP+FP
  - Recall = TP/TP+FN
  - F1 = 2* (P*R / P+R)
- BERT is quite effective
- BioBERT achieves higher scores than BERT on all the datasets
- BioBERT outperformed the state-of-the-art models on 6 out of 9 datasets

Table 6. Test results in biomedical named entity recognition

				BERT	BioBERT v1.0			BioBERT v1.1
Type	Datasets	Metrics	SOTA	(Wiki + Books)	(+ PubMed)	(+ PMC)	(+ PubMed + PMC)	(+ PubMed)
Disease	NCBI disease	P	88.30	84.12	86.76	86.16	**89.04**	88.22
		R	89.00	87.19	88.02	89.48	89.69	**91.25**
		F	88.60	85.63	87.38	87.79	89.36	**89.71**
	2010 i2b2/VA	P	87.44	84.04	85.37	85.55	**87.50**	86.93
		R	86.25	84.08	85.64	85.72	85.44	**86.53**
		F	**86.84**	84.06	85.51	85.64	86.46	86.73
	BC5CDR	P	**89.61**	81.97	85.80	84.67	85.86	86.47
		R	83.09	82.48	86.60	85.87	87.27	**87.84**
		F	86.23	82.41	86.20	85.27	86.56	**87.15**
Drug/chem.	BC5CDR	P	**94.26**	90.94	92.52	92.46	93.27	93.68
		R	92.38	91.38	92.76	92.63	**93.61**	93.26
		F	93.31	91.16	92.64	92.54	93.44	**93.47**
	BC4CHEMD	P	92.29	91.19	91.77	91.65	92.23	**92.80**
		R	90.01	88.92	90.77	90.30	90.61	**91.92**
		F	91.14	90.04	91.26	90.97	91.41	**92.36**
Gene/protein	BC2GM	P	81.81	81.17	81.72	82.86	**85.16**	84.32
		R	81.57	82.42	83.38	84.21	83.65	**85.12**
		F	81.69	81.79	82.54	83.53	84.40	**84.72**
	JNLPBA	P	**74.43**	69.57	71.11	71.17	72.68	72.24
		R	83.22	81.20	83.11	82.76	83.21	**83.56**
		F	**78.58**	74.94	76.65	76.53	77.59	77.49
Species	LINNAEUS	P	92.80	91.17	91.83	91.62	**93.84**	90.77
		R	**94.29**	84.30	84.72	85.48	86.11	85.83
		F	**93.54**	87.60	88.13	88.45	89.81	88.24
	Species-800	P	**74.34**	69.35	70.60	71.54	72.84	72.80
		R	75.96	74.05	75.75	74.71	**77.97**	75.36
		F	74.98	71.63	73.08	73.09	**75.31**	74.06

Notes: Precision (P), Recall (R) and F1 (F) scores on each dataset are reported. The best scores are in bold, and the second best scores are underlined. We list the scores of the state-of-the-art (SOTA) models on different datasets as follows: scores of Xu *et al.* (2019) on NCBI Disease, scores of Sachan *et al.* (2018) on BC2GM, scores of Zhu *et al.* (2018) (single model) on 2010 i2b2/VA, scores of Lou *et al.* (2017) on BC5CDR-disease, scores of Luo *et al.* (2018) on BC4CHEMD, scores of Yoon *et al.* (2019) on BC5CDR-chemical and JNLPBA and scores of Giorgi and Bader (2018) on LINNAEUS and Species-800.

UNIVERSITY OF TORONTO

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published. https://doi.org/10.1093/bioinformatics/btz682

BioBERT: Evaluation - RE

- BERT achieved better performance than the state-of-the-art model on the CHEMPROT dataset, which demonstrates its effectiveness in RE
- BioBERT achieved the highest F1 scores on 2 out of 3 biomedical datasets

Table 7. Biomedical relation extraction test results

Relation	Datasets	Metrics	SOTA	BERT (Wiki + Books)	BioBERT v1.0 (+ PubMed)	BioBERT v1.0 (+ PMC)	BioBERT v1.0 (+ PubMed + PMC)	BioBERT v1.1 (+ PubMed)
Gene–disease	GAD	P	**79.21**	74.28	76.43	75.20	75.95	77.32
		R	**89.25**	85.11	87.65	86.15	88.08	82.68
		F	**83.93**	79.29	81.61	80.24	81.52	79.83
	EU-ADR	P	76.43	75.45	78.04	**81.05**	80.92	77.86
		R	**98.01**	96.55	93.86	93.90	90.81	83.55
		F	85.34	84.62	84.44	**86.51**	84.83	79.74
Protein–chemical	CHEMPROT	P	74.80	76.02	76.05	**77.46**	75.20	77.02
		R	56.00	71.60	74.33	72.94	75.09	**75.90**
		F	64.10	73.74	75.18	75.13	75.14	**76.46**

Notes: Precision (P), Recall (R) and F1 (F) scores on each dataset are reported. The best scores are in bold, and the second best scores are underlined. The scores on GAD and EU-ADR were obtained from Bhasuran and Natarajan (2018), and the scores on CHEMPROT were obtained from Lim and Kang (2018).

UNIVERSITY OF TORONTO

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published. https://doi.org/10.1093/bioinformatics/btz682

BioBERT: Evaluation - QA

- All versions of BioBERT significantly outperformed BERT and the state-of-the-art models
- Strict accuracy is the rate of top 1 exact answers. Lenient accuracy is the rate of exact answers in top 5 predictions

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}.$$

Table 8. Biomedical question answering test results

			BERT	BioBERT v1.0			BioBERT v1.1
Datasets	Metrics	SOTA	(Wiki + Books)	(+ PubMed)	(+ PMC)	(+ PubMed + PMC)	(+ PubMed)
BioASQ 4b	S	20.01	27.33	25.47	26.09	**28.57**	27.95
	L	28.81	44.72	44.72	42.24	**47.82**	44.10
	M	23.52	33.77	33.28	32.42	**35.17**	34.72
BioASQ 5b	S	41.33	39.33	41.33	42.00	44.00	**46.00**
	L	56.67	52.67	55.33	54.67	56.67	**60.00**
	M	47.24	44.27	46.73	46.93	49.38	**51.64**
BioASQ 6b	S	24.22	33.54	**43.48**	41.61	40.37	42.86
	L	37.89	51.55	55.90	55.28	**57.77**	**57.77**
	M	27.84	40.88	48.11	47.02	47.48	**48.43**

Notes: Strict Accuracy (S), Lenient Accuracy (L) and Mean Reciprocal Rank (M) scores on each dataset are reported. The best scores are in bold, and the second best scores are underlined. The best BioASQ 4b/5b/6b scores were obtained from the BioASQ leaderboard (http://participants-area.bioasq.org).

UNIVERSITY OF TORONTO

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published. https://doi.org/10.1093/bioinformatics/btz682

https://en.wikipedia.org/wiki/Mean_reciprocal_rank

Agenda

- Motivation & Related Work
- Model
- Evaluation
- **Advantage & Limitation**

UNIVERSITY OF TORONTO

Advantage & Limitation

- Advantage
  - Identical framework shared between pre-training and fine-tuning process will save the cost in transfer learning
  - By providing specific training data into the fine-tuning architecture, BioBERT can be trained to solve a wide range of biomedical text mining tasks.
- Limitation
  - Limited input length leads to entity relations missing in large scale content
  - The database bias in pre-training database might affect the performance of fine-tuning

UNIVERSITY OF TORONTO

Thank You!