



CSCE 689 - Special Topics in NLP for Science

Lecture 2: Scientific LLMs (Encoder-Only & Encoder-Decoder)

Yu Zhang

yuzhang@tamu.edu

January 16, 2025

Course Website: <https://yuzhang-teaching.github.io/CSCE689-S25.html>

Sign Up for Your Presentation



- https://docs.google.com/spreadsheets/d/1sw03STrp5oyxUyXwEsd313K70L7W1Ebe_0VKzz0G2a8/edit?usp=sharing

CSCE 689 Student Presentation Sign-Up

File Edit View Insert Format Data Tools Extensions Help

Sheets home

100% | \$ % .0+ .00 123 | Defaul... | - 10 + | B I A |

	A	B	C	D
1	Week	Date	Topic	Papers
2	W3	1/30	Scientific Question Answering (2% extra credit)	* PubMedQA: A Dataset for Biomedical Research Question Answering [EMNLP 2019] * Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries [WWW 2024] * MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models [ICLR 2024]
3	W4	2/6	Paper Classification (1% extra credit)	* The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study [WWW 2023] * Hierarchical Multi-Label Classification of Scientific Documents [EMNLP 2022] * BERTMeSH: Deep Contextual Representation Learning for Large-Scale High-Performance MeSH Indexing with Full Text [Bioinformatics 2020]
4	W5	2/13	Scientific VLMs: Geometry	* UniMath: A Foundational and Multimodal Mathematical Reasoner [EMNLP 2023] * G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model [arXiv 2023] * Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models [EMNLP 2024]

Tips

- No need to cover every detail of the papers
- Focus on conveying general ideas and insights
- For the method part, **do not** go over each formula in detail, but explain the major insights
- For the experimental part, **do not** present every piece of experiment results, but explain how the empirical findings support the major claims
- Provide some take-away messages
 - Novelty, practical value, limitations, ...
- Start preparing your presentation early (e.g., 10+ days in advance)

Unable to present on your signed-up date?

- If there is a free slot, email me and I will move you to that slot.
 - Let me know at least one week before $\min\{\text{current slot}, \text{new slot}\}$.
- If you want to switch with somebody else's slot, contact the other party directly and let me know after you come to an agreement.
- The date and the papers to be presented are tied together. If you move to a new slot, you should present the corresponding new papers instead of the ones you originally signed up for.

Agenda

- Language Model Pre-training Basics
- Encoder-Only Architecture
 - BERT
 - SciBERT and BioBERT
 - ELECTRA
 - BioELECTRA
- Encoder-Decoder Architecture
 - BART and T5
 - SciFive



I will also introduce
common scientific and
biomedical NLP Tasks

Agenda

- Language Model Pre-training Basics
- Encoder-Only Architecture
 - BERT
 - SciBERT and BioBERT
 - ELECTRA
 - BioELECTRA
- Encoder-Decoder Architecture
 - BART and T5
 - SciFive

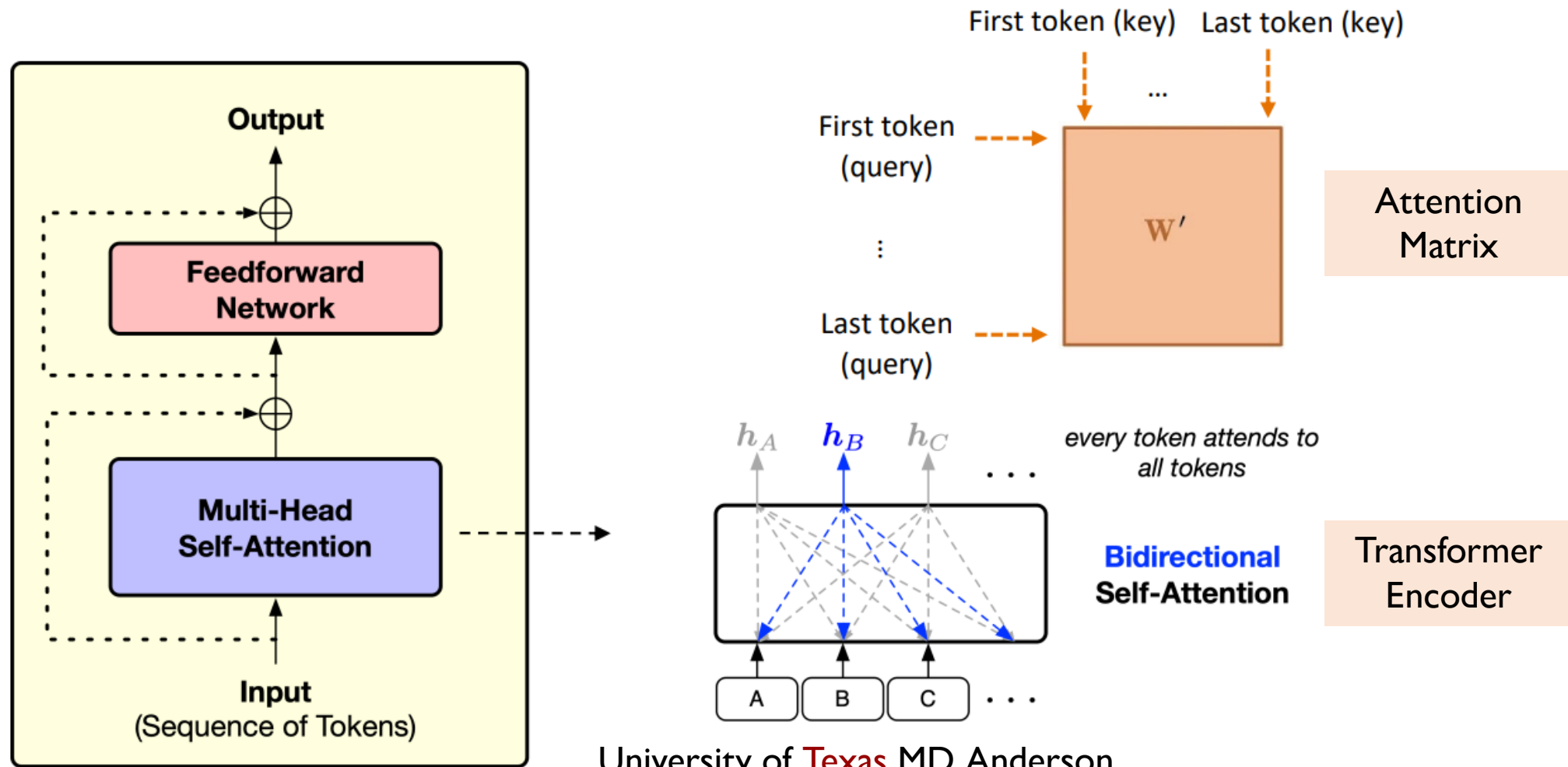
Information in Raw Texts

- *Verb:* the task is to all the structured chemical reactions from papers
 - *Preposition:* a liquid handler equipped two microplates
 - *Time:* the Transformer paper was published in
 - *Location:* data from the University of MD Anderson Cancer Center
 - *Math:* the sequence goes 1, 1, 2, 3, 5, 8, 13, 21,
 - *Chemistry:* sugar is composed of carbon, hydrogen, and
 - ...
-
- How to harvest underlying patterns, structures, and semantic knowledge from raw texts?
 - Train the model to predict masked tokens given their contexts

Information in Raw Texts

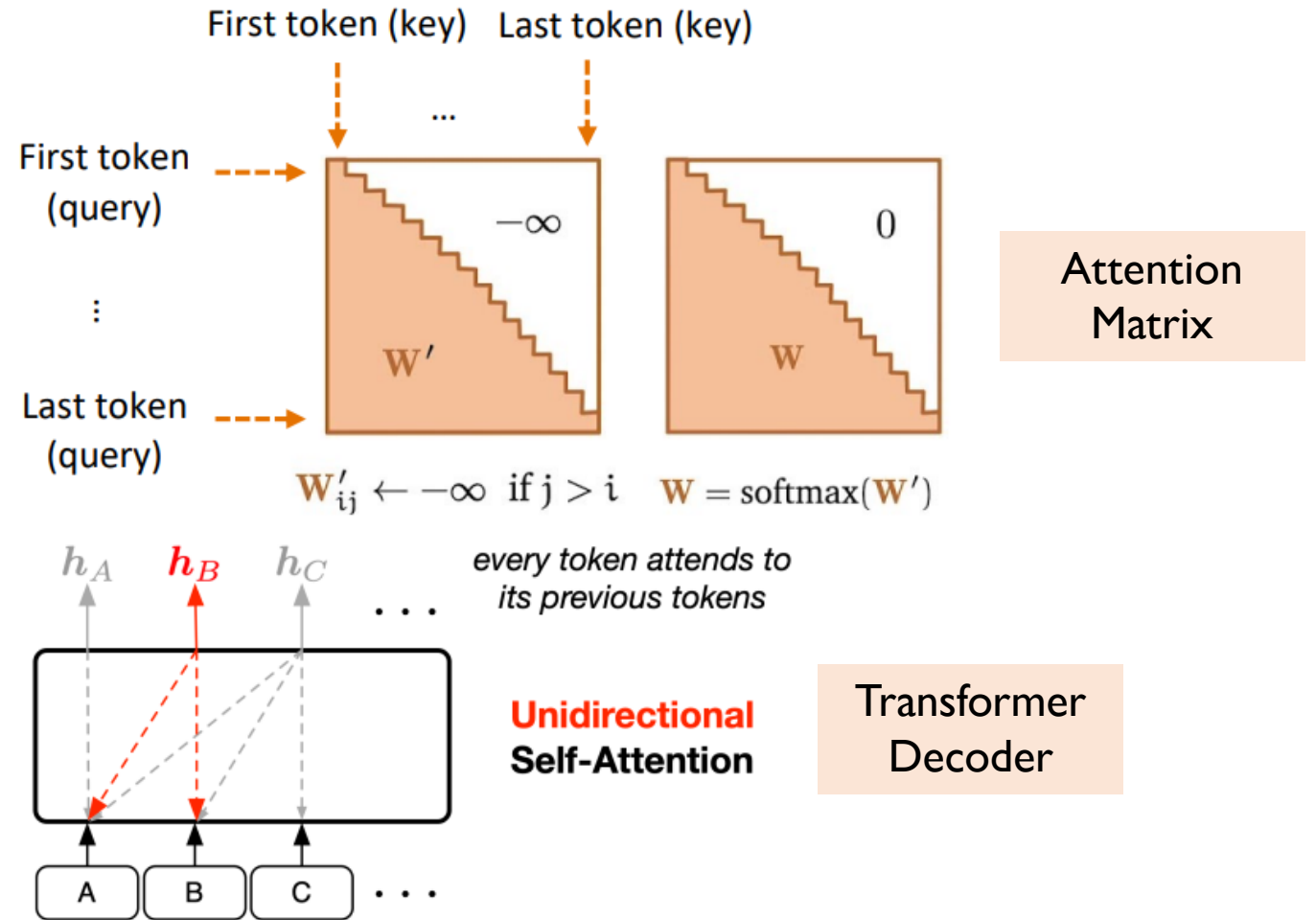
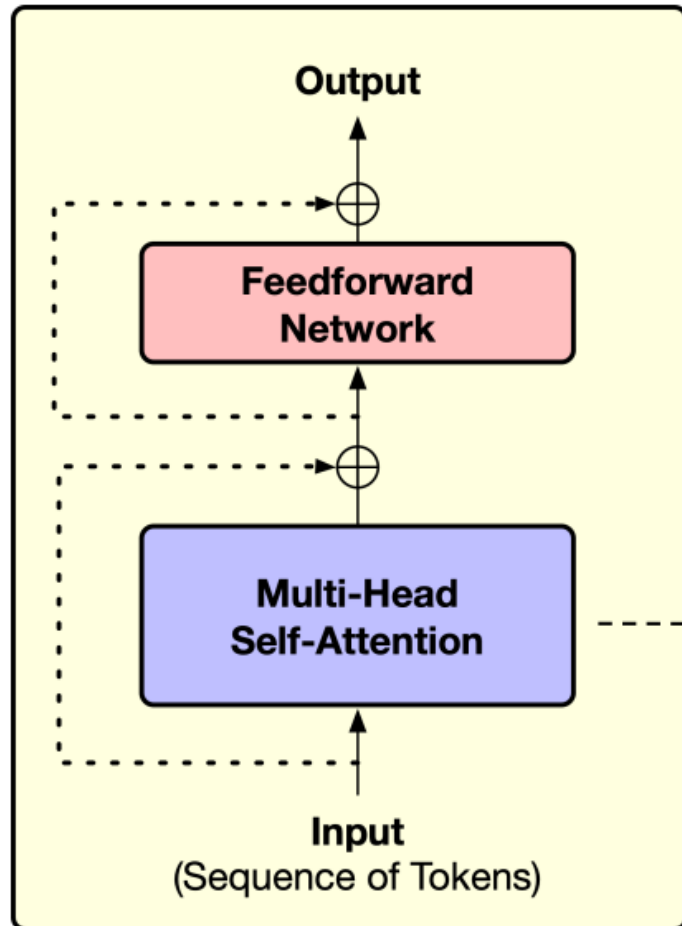
- *Verb:* the task is to extract all the structured chemical reactions from papers
 - *Preposition:* a liquid handler equipped with two microplates
 - *Time:* the Transformer paper was published in 2017
 - *Location:* data from the University of Texas MD Anderson Cancer Center
 - *Math:* the sequence goes 1, 1, 2, 3, 5, 8, 13, 21, 34
 - *Chemistry:* sugar is composed of carbon, hydrogen, and oxygen
 - ...
-
- How to harvest underlying patterns, structures, and semantic knowledge from raw texts?
 - Train the model to predict masked tokens given their contexts

The Transformer Architecture: Encoder



University of Texas MD Anderson

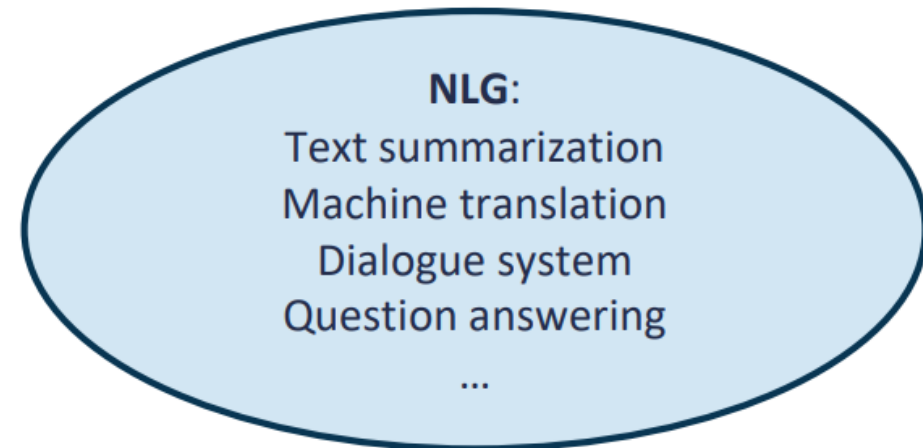
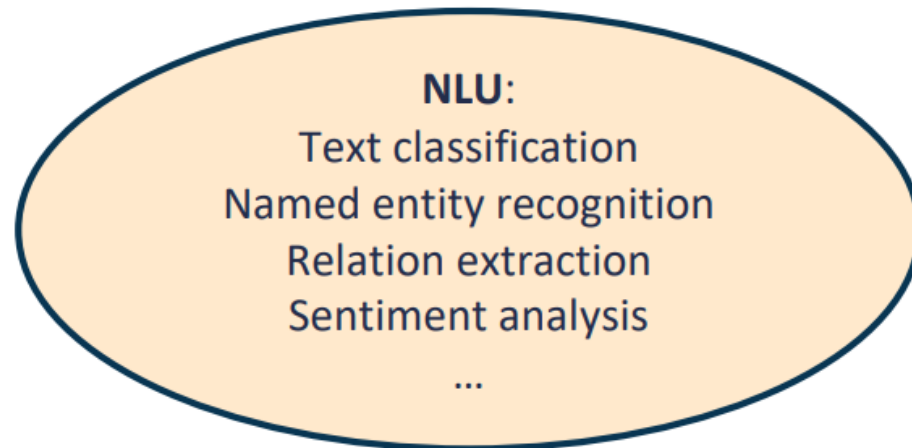
The Transformer Architecture: Decoder



published in 2017

Encoder vs. Decoder

- Encoder:
 - Each token can attend to all other tokens **to better learn its representation vector**
 - Suitable for natural language understanding (NLU) tasks
- Decoder:
 - Each token can only attend to previous tokens **to predict the next token**
 - Suitable for natural language generation (NLG) tasks

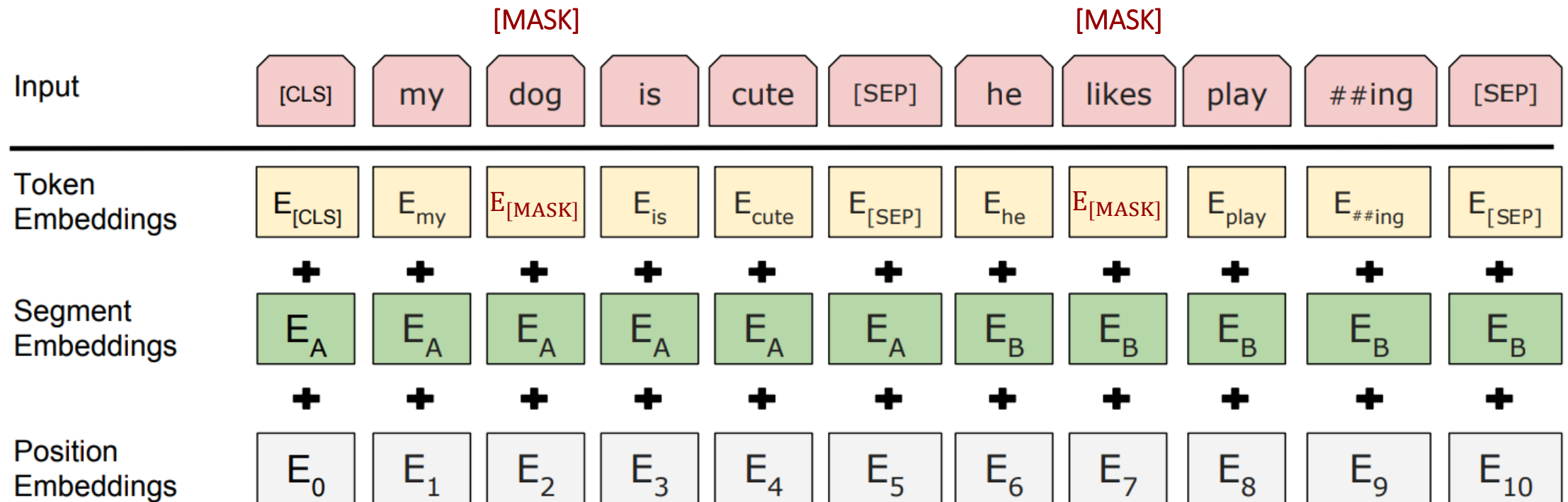


Agenda

- Language Model Pre-training Basics
- Encoder-Only Architecture
 - BERT
 - SciBERT and BioBERT
 - ELECTRA
 - BioELECTRA
- Encoder-Decoder Architecture
 - BART and T5
 - SciFive

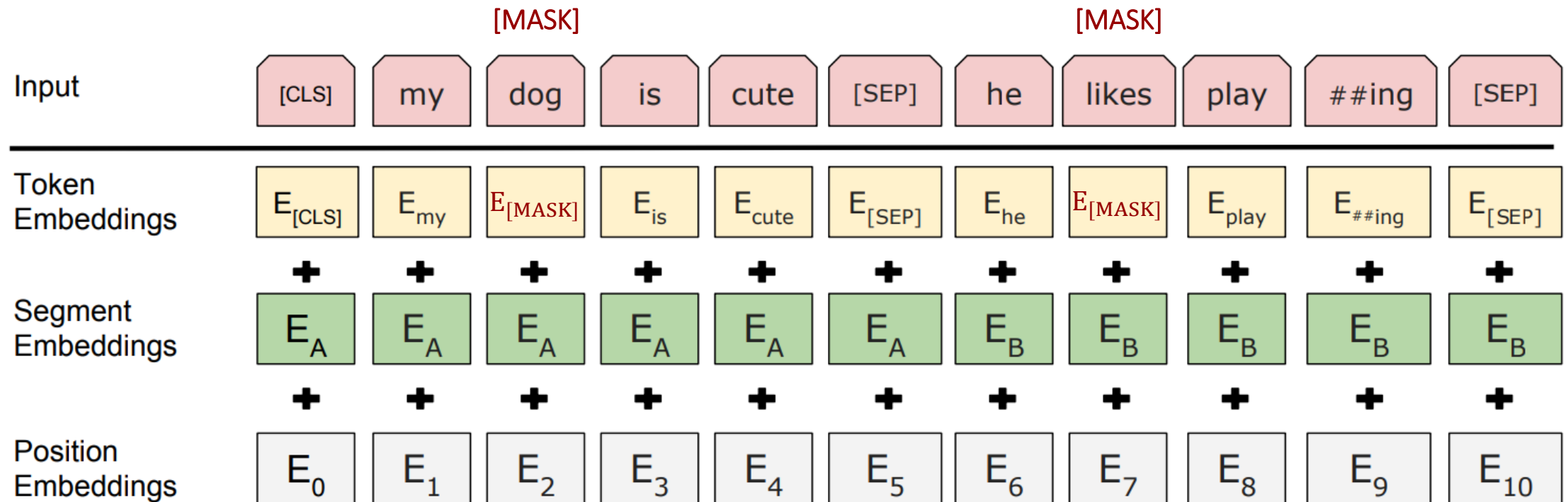
Encoder Pre-training: BERT

- **Task 1 - Masked Language Modeling (MLM):** With 15% words randomly masked, the model learns bidirectional contextual information to predict the masked words.

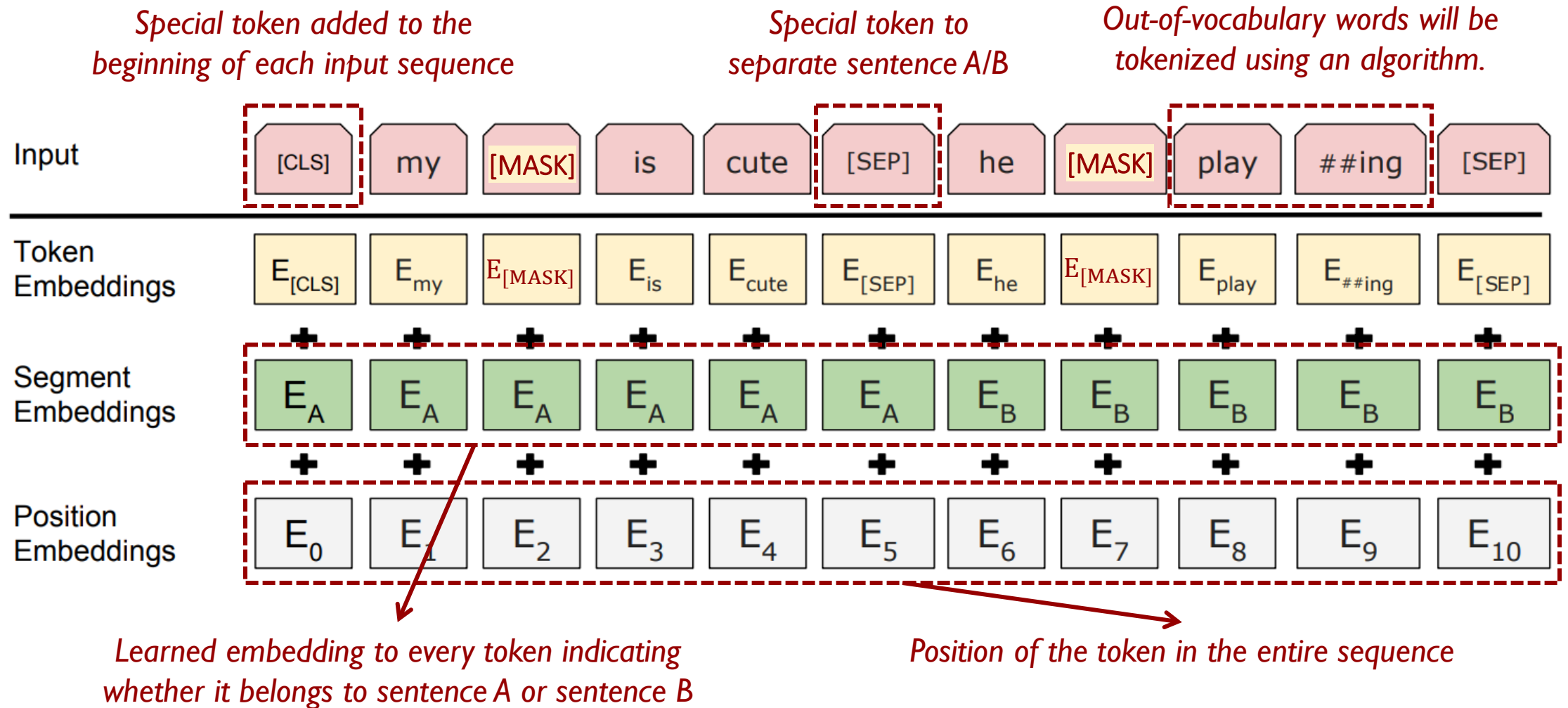


Encoder Pre-training: BERT

- **Task 2 - Next Sentence Prediction (NSP):** We feed two sentences together into the model. Predict if their order is correct. 50% of the time the two sentences are in the correct order.



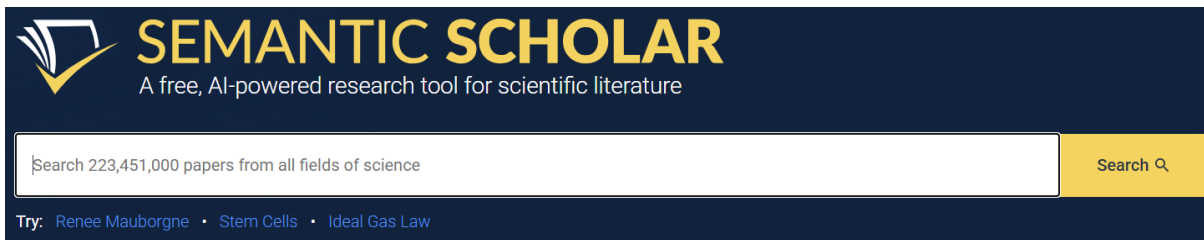
Encoder Pre-training: BERT



SciBERT

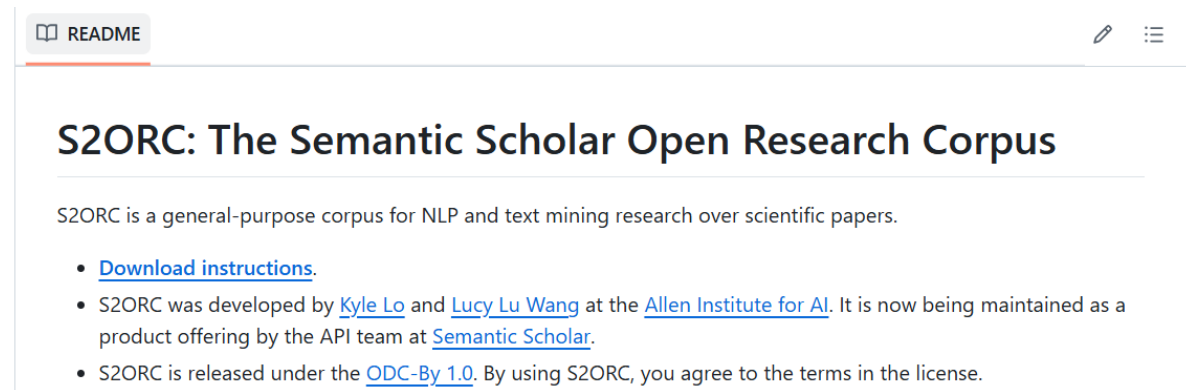
- **Architecture:** the same as BERT-base (12-layer Transformer encoders, 110M model parameters)
- **Pre-training data:** 1.14M papers from Semantic Scholar

<https://www.semanticscholar.org/>



<https://github.com/allenai/s2orc>

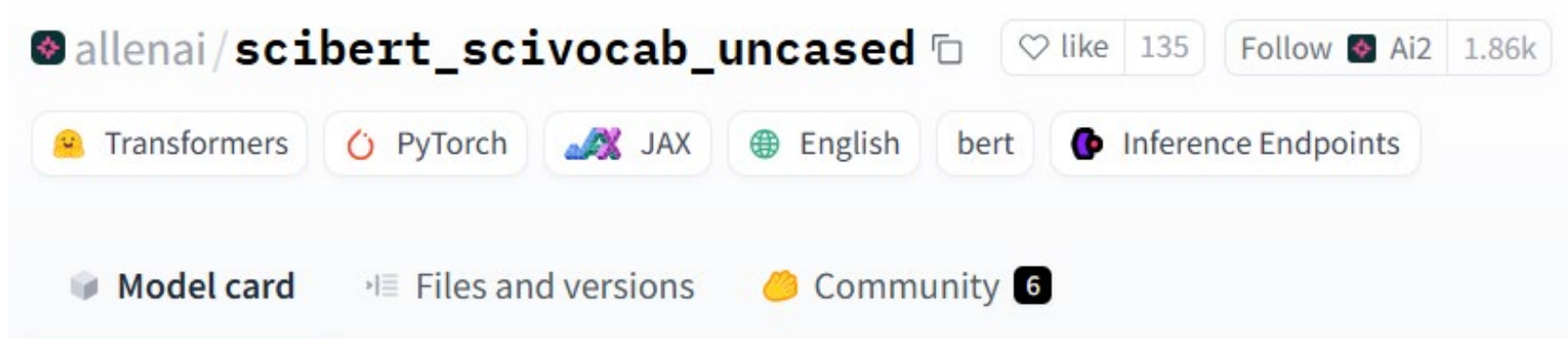
title, abstract, full text, ...



SciBERT

- Model variants:
 - *Uncased* or *Cased*: *Uncased* performs slightly better
 - *From BERT* or *From Scratch*: *From Scratch* performs slightly better because the model can adopt a domain-specific vocabulary rather than stick with the BERT vocabulary

https://huggingface.co/allenai/scibert_scivocab_uncased



Scientific NLP Tasks for Evaluating SciBERT: NER

- **Named Entity Recognition (NER):** Given a sentence, find entities (i.e., token spans) of certain types (e.g., chemical, disease, gene).

... human complement factor H deficiency associated with hemolytic uremic syndrome ...

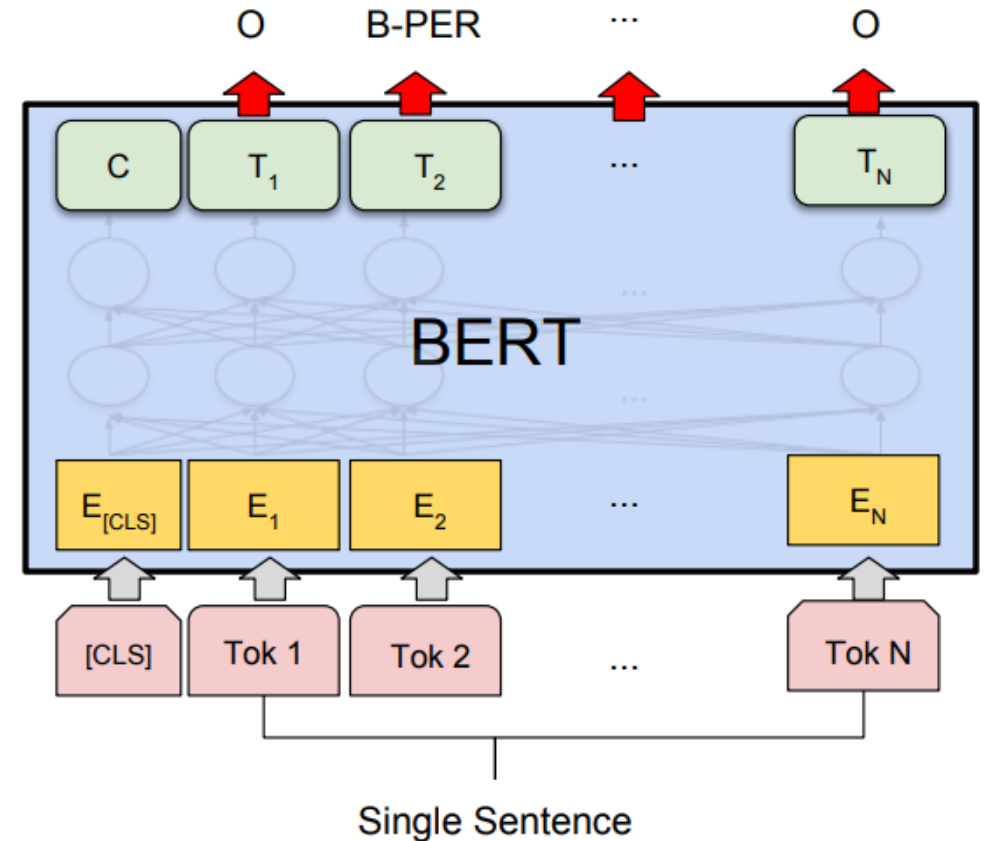
DISEASE DISEASE

Input	human	complement	factor	H	deficiency	associated	with
Output	B-DISEASE	I-DISEASE	I-DISEASE	I-DISEASE	I-DISEASE	O	O

- The BIO schema: B (beginning of an entity), I (in an entity), O (out of an entity)
 - Why do we need “B” rather than just have “I” and “O”?
- NER → predicting a label for each token in the sentence

Fine-tuning the BERT Architecture for NER

- Evaluation datasets used in SciBERT:
 - Biomedicine
 - [BC5CDR](#) (Chemical, Disease)
 - [JNLPBA](#) (Gene/Protein, DNA, Cell Type, Cell Line, RNA)
 - [NCBI-disease](#) (Disease)
 - Computer Science
 - [SciERC](#) (Task, Method, Metric, Material)



Scientific NLP Tasks for Evaluating SciBERT: Classification

- **Topic Classification:** Given a paper, predict its topic (e.g., natural language processing vs. computer vision).
- **Citation Intent Classification:** Given a citation sentence in a paper, predict the intent of this citation.

Class	Description	Example
BACKGROUND	P provides relevant information for this domain.	This is often referred to as incorporating deterministic closure (Dörre, 1993).
MOTIVATION	P illustrates need for data, goals, methods, etc.	As shown in Meurers (1994), this is a well-motivated convention [...]
USES	Uses data, methods, etc., from P .	The head words can be automatically extracted [...] in the manner described by Magerman (1994).
EXTENSION	Extends P 's data, methods, etc.	[...] we improve a two-dimensional multimodal version of LDA (Andrews et al, 2009) [...]
COMPARISON OR CONTRAST	Expresses similarity/differences to P .	Other approaches use less deep linguistic resources (e.g., POS-tags Stymne (2008)) [...]
FUTURE	P is a potential avenue for future work.	[...] but we plan to do so in the near future using the algorithm of Littlestone and Warmuth (1992).

Scientific NLP Tasks for Evaluating SciBERT: Classification

- **Topic Classification:** Given a paper, predict its topic (e.g., natural language processing vs. computer vision).
- **Citation Intent Classification:** Given a citation sentence in a paper, predict the intent of this citation.
- **Relation Classification (a.k.a., Relation Extraction, RE):** Given a sentence containing two entities, predict the relation between these two entities.

In rats, Nitrofurantoin causes pulmonary toxicity.

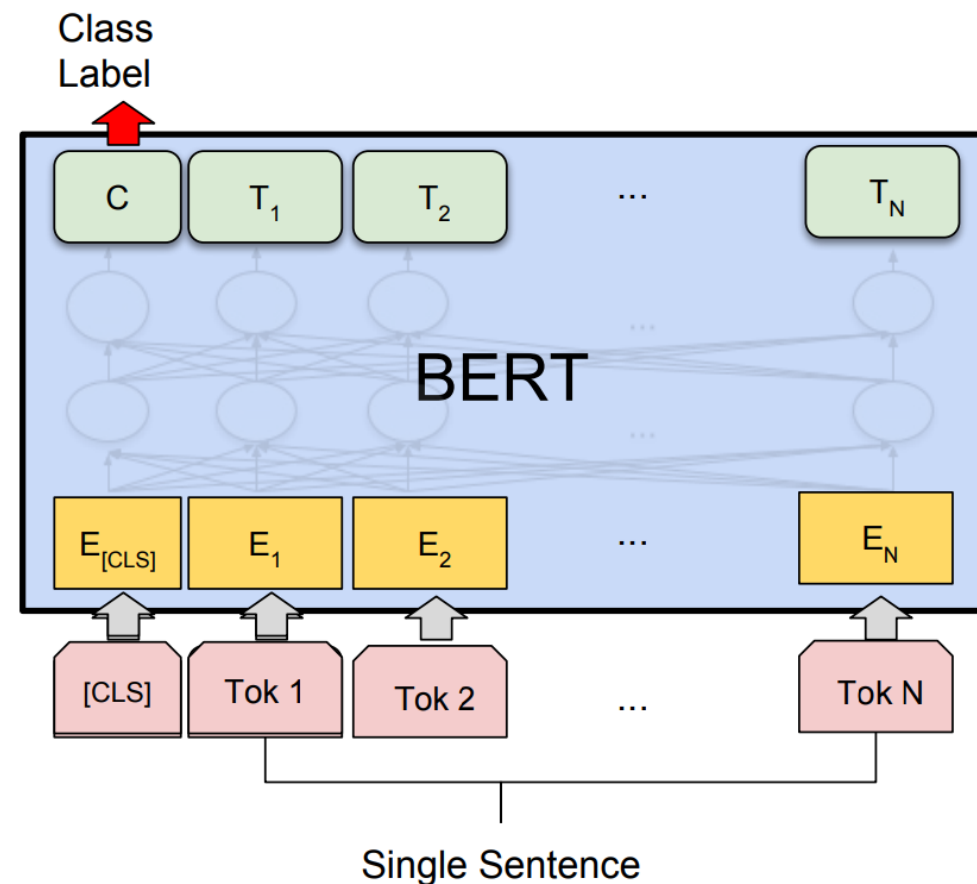
CHEMICAL

DISEASE

Predicted Relation: CHEMICAL induce DISEASE ✓
CHEMICAL treat DISEASE ✗

Fine-tuning the BERT Architecture for Classification

- Evaluation datasets used in SciBERT
 - [Microsoft Academic](#) (Paper Classification: economics vs. business vs. sociology vs. ...)
 - [ACL-ARC](#) and [SciCite](#) (Citation Intent Classification)
 - [ChemProt](#) (Relation Classification: Protein-Chemical)
 - [SciERC](#) (Relation Classification: Task-Task, Method-Task, ...)
- Better ways than using [CLS] alone?
 - Average of all tokens
 - Concatenating entity representations



BioBERT

- **Architecture:** the same as BERT-base (12-layer Transformer encoders, 110M model parameters) and BERT-large (24-layer Transformer encoders, 340M model parameters)
- **Pre-training data:** PubMed abstracts (~30M papers, 4.5B words) + PMC full text (13.5B words)

<https://pubmed.ncbi.nlm.nih.gov>



<https://pmc.ncbi.nlm.nih.gov>



<https://ftp.ncbi.nlm.nih.gov/pubmed/baseline>




Index of /pubmed/baseline





Name	Last modified	Size
Parent Directory		-
README.txt	2024-01-12 14:25	4.5K
pubmed24n0001.xml.gz	2023-12-14 18:08	19M
pubmed24n0001.xml.gz.md5	2023-12-14 18:08	60
pubmed24n0002.xml.gz	2023-12-14 18:08	17M
pubmed24n0002.xml.gz.md5	2023-12-14 18:08	60
pubmed24n0003.xml.gz	2023-12-14 18:08	16M
pubmed24n0003.xml.gz.md5	2023-12-14 18:08	60




BioBERT

- Model variants:
 - *Uncased* or *Cased*: *Cased* performs slightly better (different from SciBERT!)
 - *PubMed* or *PMC* or *PubMed+PMC*: More data results in better performance
 - *200K Steps* or *1M Steps* in pre-training: Longer pre-training results in better performance

<https://huggingface.co/dmis-lab/biobert-v1.1>

 dmis-lab / **biobert-v1.1**   like 72

 Feature Extraction  Transformers  PyTorch  JAX

 Model card  Files and versions  Community **6**

<https://huggingface.co/dmis-lab/biobert-large-cased-v1.1>

 dmis-lab / **biobert-large-cased-v1.1**   like 5

 Transformers  PyTorch  Inference Endpoints

 Model card  Files and versions  Community **1**

Scientific NLP Tasks for Evaluating BioBERT: NER and RE

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

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

Scientific NLP Tasks for Evaluating BioBERT: QA

- **(Extractive) Question Answering:** Given a context paragraph and a question, extract the answer (a span of tokens) from the context graph.

Context: Corynebacterium minutissimum is the bacteria that leads to cutaneous eruptions of erythrasma and is the most common cause of interdigital foot infections. It is found mostly in occluded intertriginous areas such as the axillae, ...

Question: Which bacteria causes erythrasma?

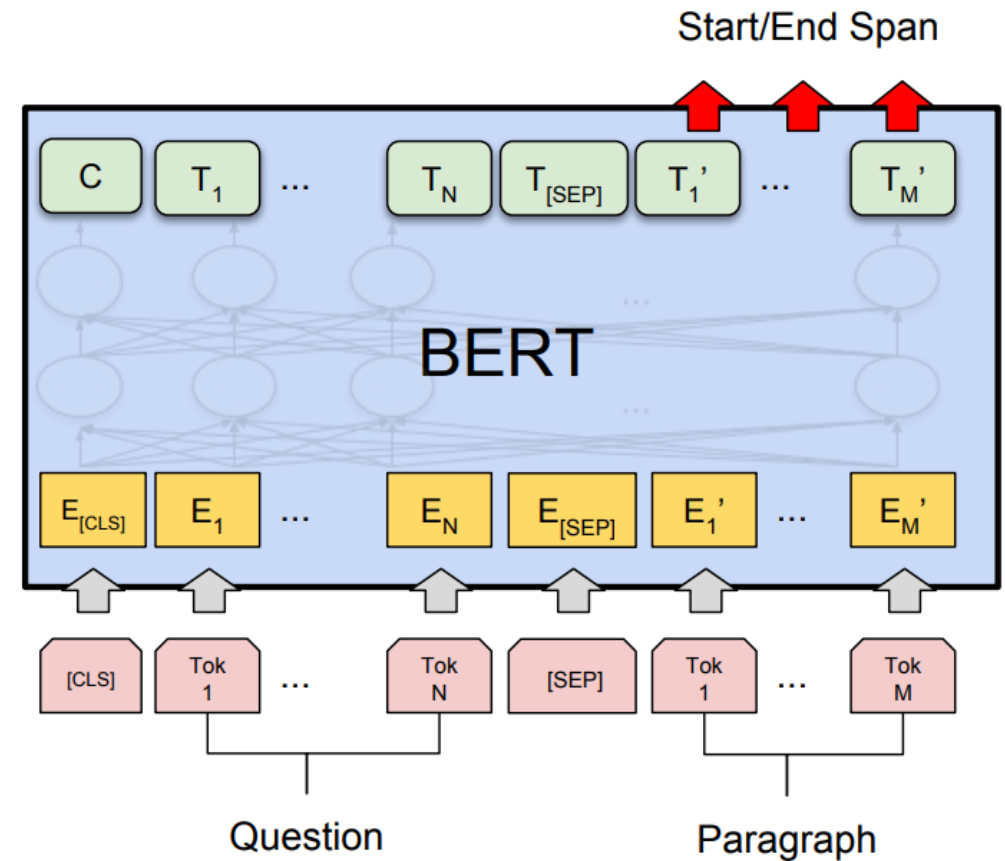


Answer: Corynebacterium minutissimum is the bacteria that leads to cutaneous ...

Fine-tuning the BERT Architecture for QA

- Evaluation datasets used in BioBERT

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



Experimental Results of SciBERT

Field	Task	Dataset	SOTA	BERT-Base		SciBERT	
				Frozen	Finetune	Frozen	Finetune
Bio	NER	BC5CDR (Li et al., 2016)	88.85 ⁷	85.08	86.72	88.73	90.01
		JNLPBA (Collier and Kim, 2004)	78.58	74.05	76.09	75.77	77.28
		NCBI-disease (Dogan et al., 2014)	89.36	84.06	86.88	86.39	88.57
	PICO	EBM-NLP (Nye et al., 2018)	66.30	61.44	71.53	68.30	72.28
	DEP	GENIA (Kim et al., 2003) - LAS	91.92	90.22	90.33	90.36	90.43
		GENIA (Kim et al., 2003) - UAS	92.84	91.84	91.89	92.00	91.99
	REL	ChemProt (Kringelum et al., 2016)	76.68	68.21	79.14	75.03	83.64
CS	NER	SciERC (Luan et al., 2018)	64.20	63.58	65.24	65.77	67.57
	REL	SciERC (Luan et al., 2018)	n/a	72.74	78.71	75.25	79.97
	CLS	ACL-ARC (Jurgens et al., 2018)	67.9	62.04	63.91	60.74	70.98
Multi	CLS	Paper Field	n/a	63.64	65.37	64.38	65.71
		SciCite (Cohan et al., 2019)	84.0	84.31	84.85	85.42	85.49
Average				73.58	77.16	76.01	79.27

Experimental Results of BioBERT (NER)

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

Experimental Results of BioBERT (RE and QA)

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

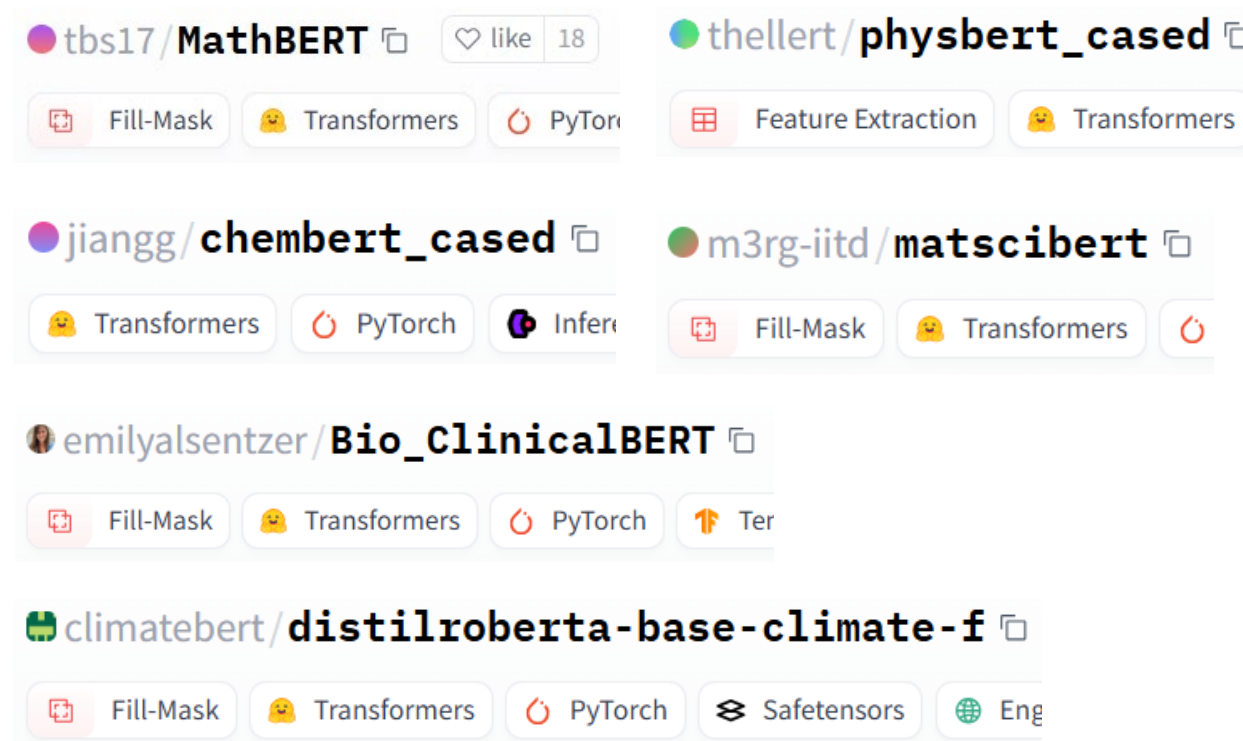
Datasets	Metrics	SOTA	BERT	BioBERT v1.0			BioBERT v1.1
			(Wiki + Books)	(+ PubMed)	(+ PMC)	(+ PubMed + PMC)	(+ PubMed)
BioASQ 4b	S	20.01	27.33	25.47	26.09	28.57	<u>27.95</u>
	L	28.81	<u>44.72</u>	<u>44.72</u>	42.24	47.82	44.10
	M	23.52	33.77	33.28	32.42	35.17	<u>34.72</u>
BioASQ 5b	S	41.33	39.33	41.33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

Take-Away Messages from SciBERT and BioBERT

- Various scientific NLP tasks can be solved by fine-tuning a pre-trained scientific language model.
- Pre-trained language models, although task-agnostic during pre-training, achieve competitive (and often the best) performance in comparison with previous task-specific architectures.
- Domain-specific BERT outperforms general BERT in domain-specific tasks.
- More pre-training data (within the domain) often helps.
- Longer pre-training steps often helps.

BERT in Other Scientific Fields

- Mathematics: MathBERT [1]
- Physics: PhysBERT [2]
- Chemistry: ChemBERT [3]
- Materials Science: MatSciBERT [4]
- Medicine: ClinicalBERT [5]
- Geoscience: ClimateBERT [6]



- [1] *MathBERT: A Pre-trained Language Model for General NLP Tasks in Mathematics Education*. arXiv 2021
- [2] *PhysBERT: A Text Embedding Model for Physics Scientific Literature*. arXiv 2024
- [3] *Automated Chemical Reaction Extraction from Scientific Literature*. Journal of Chemical Information and Modeling 2022
- [4] *MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction*. npj Computational Materials 2022
- [5] *Publicly Available Clinical BERT Embeddings*. NAACL 2019 Workshop
- [6] *ClimateBERT: A Pretrained Language Model for Climate-Related Text*. arXiv 2021

Improving BERT: RoBERTa

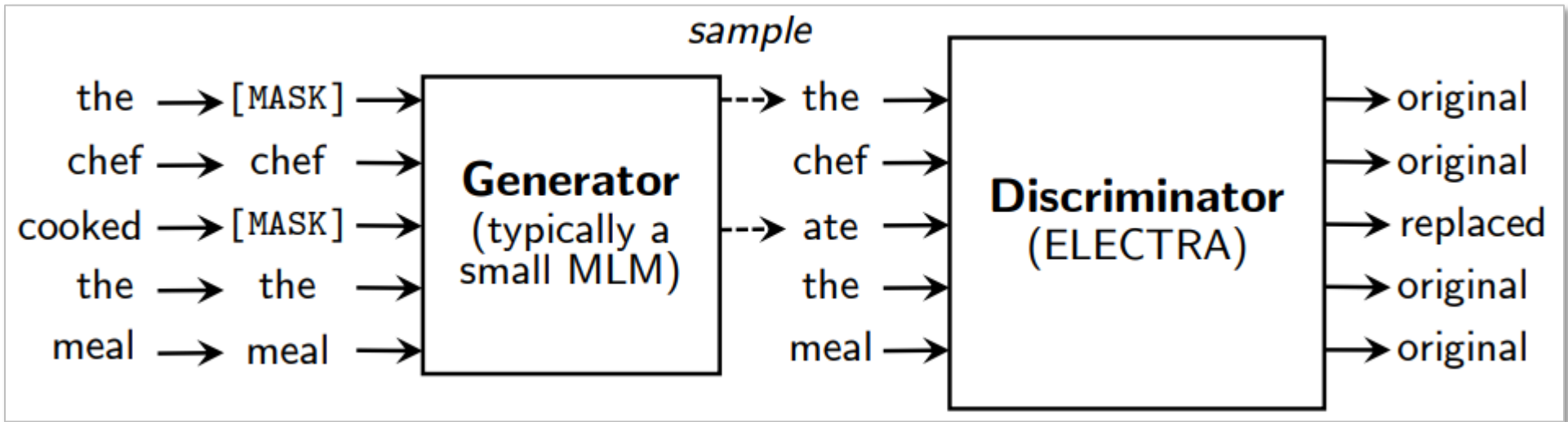
- BERT has two pre-training tasks: masked language modeling (MLM) and next sentence prediction (NSP).
- Next sentence prediction is not helpful!
- Pretrain on longer sequences
- Pretrain the model for longer, with bigger batches over more data
- Dynamically change the masking patterns applied to the training data in each epoch

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
<i>Our reimplementation (with NSP loss):</i>				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
<i>Our reimplementation (without NSP loss):</i>				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	94.6/89.4	90.2	96.4

Improving BERT: ELECTRA

- Use a small MLM as an auxiliary generator (discarded after pretraining)
- Pretrain the main model as a discriminator
- The small auxiliary MLM and the main discriminator are jointly trained.
- The main model's pretraining task becomes **more and more challenging** in pretraining.



ELECTRAMed

- **Architecture:** the discriminator is the same as BERT-base (12-layer Transformer encoders, 110M model parameters); the generator is $\frac{1}{4}$ the size of the discriminator in terms of hidden dimensions
- **Pre-training data:** PubMed abstracts

https://huggingface.co/giacomomiolo/electrased_base_scivocab_1M



Evaluating ELECTRAMed (NER and RE)

- **NER:** NCBI-disease, BC5CDR, and JNLPBA

Dataset	SciBERT	BioBERT	ELECTRAMed
NCBI-disease	89.36	89.71	87.54
BC5CDR	88.85	-	90.03
JNLPBA	78.58	-	73.65

- **RE:** ChemProt and DDI-2013

Dataset	SciBERT	BioBERT	ELECTRAMed
ChemProt	83.64	76.46	72.94
DDI-2013	-	-	79.13

Evaluating ELECTRAMed (QA)

- **QA**: Instead of trying previous benchmark datasets, the authors of ELECTRAMed attended that year's BioASQ challenge (7b).

<https://www.bioasq.org>



A challenge on large-scale
biomedical semantic indexing
and question answering

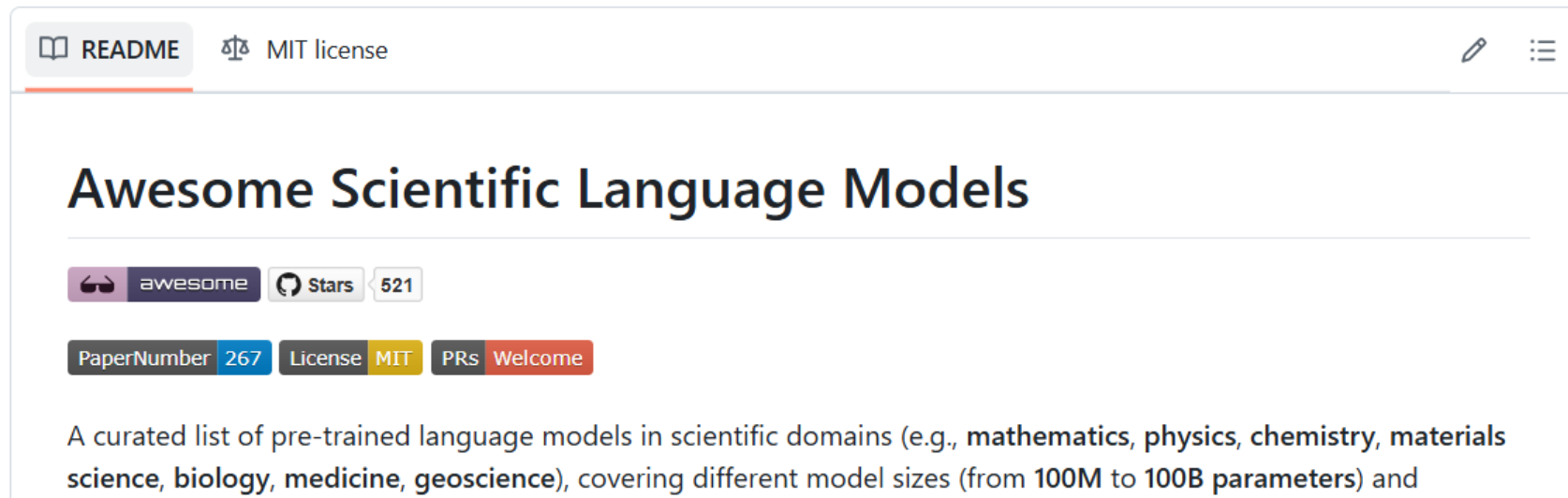


Batch	Competitor	SACC	LACC	MRR
1	(1) ELECTRAMed	44.62	51.28	<u>47.95</u>
	(2) KU-DMIS-1	41.03	53.85	46.37
	(3) BJUTNLPGroup	30.77	41.03	34.83
	(4) auth-qa-1	25.64	30.77	27.78
2	(1) KU-DMIS-5	52.00	64.00	<u>56.67</u>
	(2) ELECTRAMed	46.40	62.40	53.16
	(3) QA1	36.00	48.00	40.33
	(4) transfer-learning	24.00	44.00	32.67
3	(1) QA1	44.83	58.62	<u>51.15</u>
	(2) UNCC_QA_1	44.83	58.62	51.15
	(3) google-gold-input	41.38	65.52	50.23
	(7) ELECTRAMed	37.93	58.62	46.62
4	(1) ELECTRAMed	61.18	82.35	<u>69.55</u>
	(2) KU-DMIS-1	58.82	82.35	69.12
	(3) FACTOIDS	52.94	73.53	61.03
	(4) UNCC_QA3	52.94	73.53	61.03
5	(1) KU-DMIS-5	28.57	51.43	<u>36.38</u>
	(2) BJUTNLPGroup	28.57	40.00	33.81
	(3) UNCC_QA_1	28.57	42.86	33.05
	(6) ELECTRAMed	24.57	44.00	31.42

Take-Away Messages from ELECTRAMed

- The success of new pre-training techniques may not be generalized well to certain scientific domains.
- Most studies on encoder-based scientific LLMs still adopt the BERT architecture and its MLM pre-training objective.

<https://github.com/yuzhimanhua/Awesome-Scientific-Language-Models>

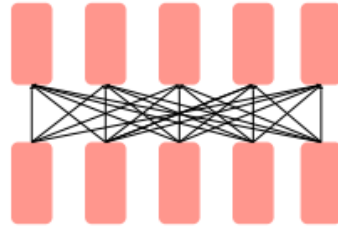


Agenda

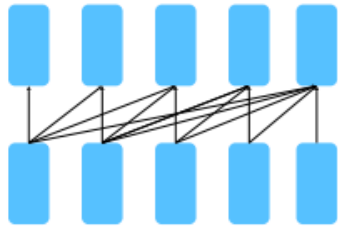
- Language Model Pre-training Basics
- Encoder-Only Architecture
 - BERT
 - SciBERT and BioBERT
 - ELECTRA
 - BioELECTRA
- Encoder-Decoder Architecture
 - BART and T5
 - SciFive

Encoder-Decoder Architecture

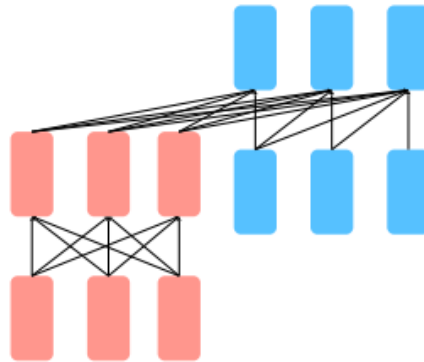
Encoder



Decoder



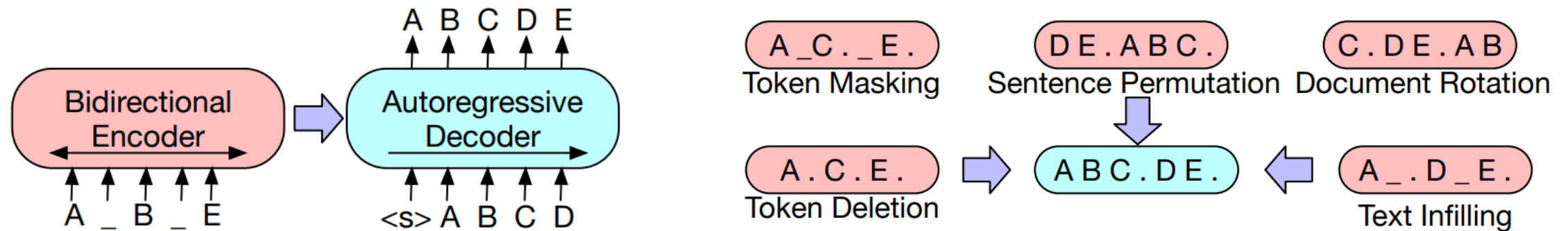
Encoder-Decoder



- What are the advantages of Encoder-Decoder architectures?
 - Can perform both NLU and NLG tasks
 - If you input “Translate *Bank of America* into Spanish” into an LLM, ...
 - What is the context of *Bank* if you use a Decoder-Only architecture?
 - What if you use an Encoder-Decoder?

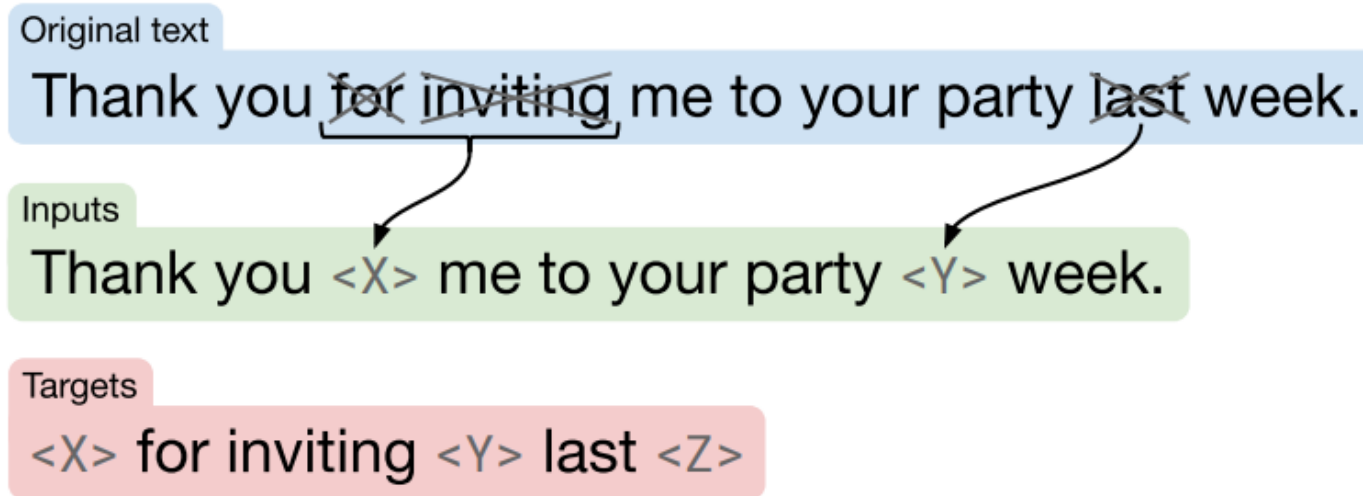
BART

- **Pre-training:** Apply a series of noising schemes (e.g., masks, deletions, permutations, etc.) to input sequences and train the model to recover the original sequences
- **Fine-tuning:**
 - For NLU tasks: Feed the same input into the encoder and decoder, and use the final decoder token for classification
 - For NLG tasks: The encoder takes the input sequence, and the decoder generates outputs autoregressively



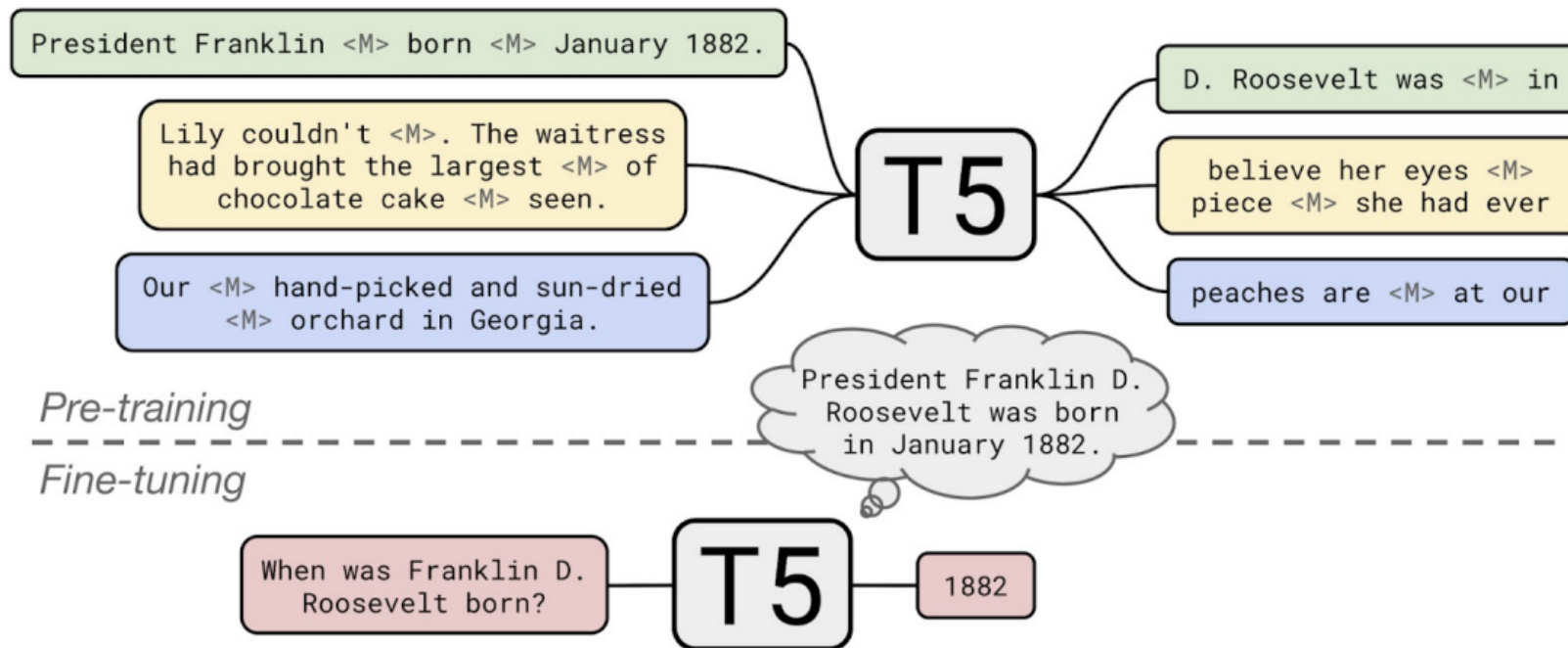
T5

- T5: Text-to-Text Transfer Transformer
- Pre-training: Mask out spans of texts; generate the original spans



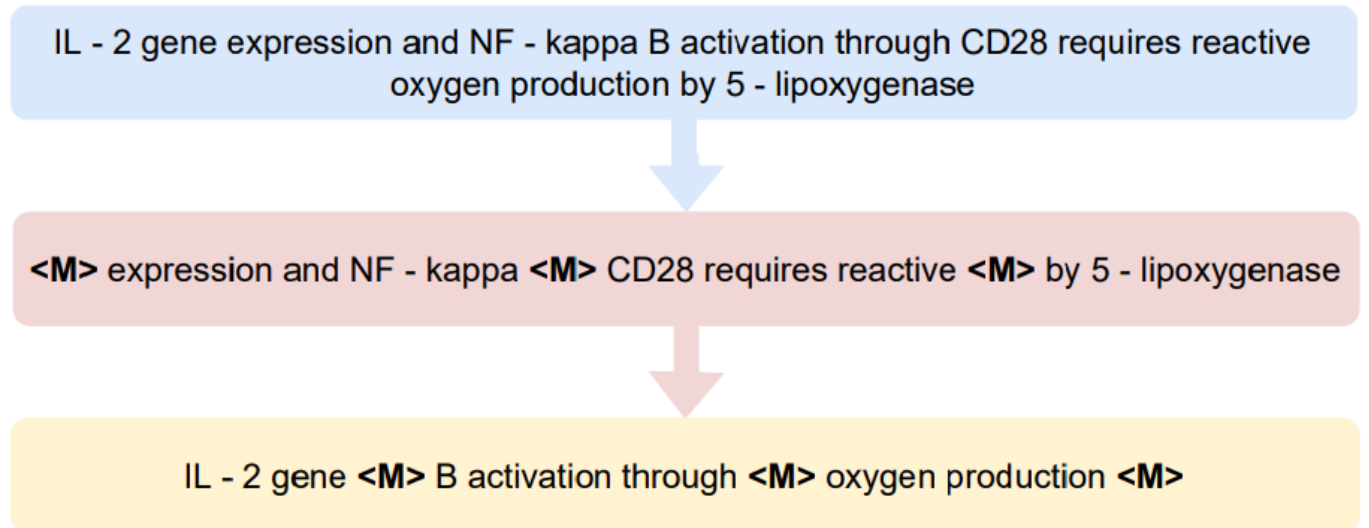
T5

- T5: Text-to-Text Transfer Transformer
- **Pre-training:** Mask out spans of texts; generate the original spans
- **Fine-tuning:** Convert every task into a sequence-to-sequence generation problem



SciFive

- **Architecture:**
 - T5-base: 12-layer Transformer encoders, 12-layer Transformer decoders, 220M parameters
 - T5-large: 24-layer Transformer encoders, 24-layer Transformer decoders, 770M parameters
- **Pre-training data:** PubMed abstracts + PMC full text



Evaluating SciFive

			Metrics	SOTA	Bert (base)	T5	BlueBERT	BioBert	SciFive (PMC +PubMed)	SciFive (PubMed)	SciFive (PMC)	T5	BlueBERT	BioBert	SciFive (PMC +PubMed)	SciFive (PubMed)	SciFive (PMC)
NER	Disease	NCBI disease	P		84.12	87.18	-	<u>88.22</u>	88.28	86.28	88.65	87.48	-	87.70	88.10	88.52	87.64
			R		87.19	89.93	-	91.25	89.30	89.71	<u>90.14</u>	90.14	-	89.90	90.14	89.82	89.30
			F	88.60	85.63	88.54	-	89.71	88.79	87.96	<u>89.39</u>	88.78	-	88.79	89.11	89.17	88.46
		BC5CDR Disease	P		81.97	85.95	-	86.47	86.67	86.53	86.48	84.28	-	-	<u>86.73</u>	86.30	87.01
			R		82.48	87.73	-	87.84	88.01	88.37	87.99	87.38	-	-	<u>88.46</u>	87.67	<u>88.24</u>
			F	86.23	82.41	86.83	86.6	87.15	87.33	<u>87.44</u>	87.23	86.31	83.8	-	87.59	86.98	87.62
	Drug/chem	BC5CDR Chemical	P		90.94	93.30	-	93.68	93.89	94.01	<u>94.09</u>	93.44	-	93.18	94.13	93.98	93.86
			R		91.38	93.92	-	93.26	94.80	94.69	94.28	95.02	-	92.09	95.39	95.36	<u>95.37</u>
			F	93.31	91.16	93.61	93.5	93.47	94.34	94.35	94.18	94.22	91.7	92.63	94.76	<u>94.66</u>	94.61
		BC4CHEMD	P		91.19	90.57	-	92.80	92.50	92.71	92.01	91.19	-	93.00	<u>92.89</u>	92.19	91.98
			R		88.92	88.90	-	<u>91.92</u>	91.53	91.35	91.87	88.76	-	92.35	91.17	91.73	91.15
			F	91.14	90.04	89.73	-	<u>92.36</u>	92.01	92.02	92.07	89.96	-	92.67	92.03	91.96	91.56
	Gene/protein	BC2GM	P		81.17	82.43	-	84.32	84.44	84.97	83.66	82.63	-	<u>84.78</u>	84.20	83.81	83.95
			R		82.42	82.17	-	<u>85.12</u>	83.89	82.89	83.04	82.10	-	85.25	83.48	83.39	83.20
			F	81.69	81.79	82.29	-	<u>84.72</u>	84.16	83.92	84.29	82.36	-	85.01	83.84	83.60	83.57
		JNLPBA	P		69.57	69.35	-	<u>72.24</u>	70.36	70.91	70.65	71.04	-	-	71.08	71.36	77.68
			R		81.20	80.61	-	83.56	80.96	80.96	81.99	81.31	-	-	<u>81.62</u>	81.46	77.42
			F	78.58	74.94	74.56	-	77.49	75.29	75.60	75.89	75.83	-	-	75.99	76.08	77.55
	SPECIES	Species-800	P		69.35	72.18	-	72.80	73.47	<u>73.84</u>	72.68	72.69	-	-	72.55	73.08	74.09
			R		74.05	76.59	-	75.36	<u>79.33</u>	79.45	79.83	76.84	-	-	77.33	78.08	78.71
			F	74.98	71.63	74.32	-	74.06	76.29	76.55	76.08	74.66	-	-	74.86	75.50	<u>76.33</u>
RE	ChemProt		P	74.80	76.02	81		77.02	82.59	84.24	82.35	<u>84.04</u>	-	-	81.99	81.31	83.58
			R	56.00	71.60	89.01		75.90	91.21	93.96	92.31	86.81	-	-	95.06	95.60	<u>95.06</u>
			F	64.10	73.74	84.82	72.5	76.46	86.68	<u>88.83</u>	87.04	85.41	74.4	-	88.04	87.88	88.95
	DDI		P	-	-	82.68	-	-	81.96	83.15	82.75	83.87	-	-	84.22	<u>83.88</u>	83.00
			R	-	-	81.41	-	-	83.04	83.15	82.33	82.84	-	-	82.84	<u>83.45</u>	84.27
			F	72.9	-	82.04	79.4	-	82.50	83.15	82.54	83.35	79.9	-	83.52	83.67	<u>83.63</u>
DoC	HoC		P	-	-	85.55	-	-	86.27	86.18	86.08	86.02	-	-	86.11	86.35	86.36
			R	-	-	85.42	-	-	86.29	86.17	86.20	85.95	-	-	86.21	86.31	86.39
			F*	81.5	-	85.22	85.3	-	85.99	85.89	85.83	85.68	87.3	-	85.87	86.03	<u>86.08</u>
NLI	MedNLI		Acc	73.5	-	83.90	84.0	-	84.88	85.30	84.25	83.8	83.8	-	86.57	<u>86.36</u>	86.08

The Natural Language Inference (NLI) Task

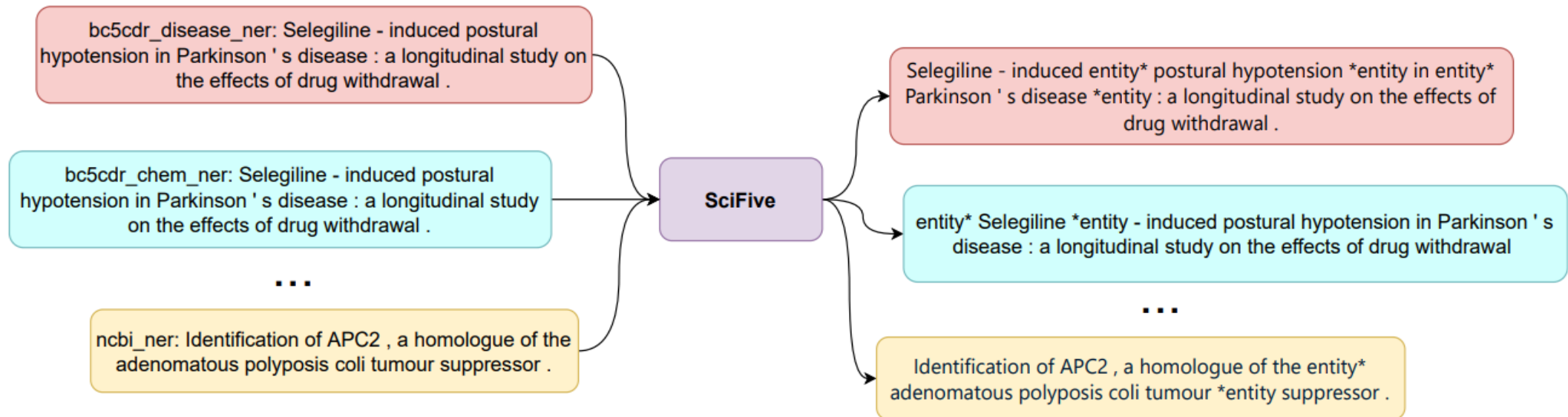
- Given two sentences (a premise and a hypothesis), predict whether the premise entails the hypothesis.

Premise	Hypothesis	Label
The patient presents with a high fever, persistent cough, and shortness of breath.	The patient may have a respiratory infection.	Entailment
The patient presents with a high fever, persistent cough, and shortness of breath.	The patient does not have symptoms.	Contradiction
The patient presents with a high fever, persistent cough, and shortness of breath.	The patient is recovering from surgery.	Neutral

- Intuitively, next sentence prediction should help NLI.
 - For two adjacent sentences, the first often entails the second.

Fine-tuning SciFive

- Convert every task into a sequence-to-sequence generation problem
- Taking NER as an example:



- What if you have a new NER task (e.g., cell type NER) without any training data?
 - Next lecture!



Thank You!

Course Website: <https://yuzhang-teaching.github.io/CSCE689-S25.html>